CCL 2024

# The 23rd Chinese National Conference on Computational Linguistics

July 24 - July 28, 2024

Taiyuan, China

©The 23rd Chinese National Conference on Computational Linguistics

Order copies of this and other CCL proceedings from:

Chinese National Conference on Computational Linguistics (CCL)

Courtyard 4, South Fourth Street, Zhongguancun , Haidian District, Beijing 100190, China

Tel: + 010-62562916

Fax: + 010-62661046

cips@iscas.ac.cn

# Introduction

Welcome to the proceedings of the twenty third China National Conference on Computational Linguistics (23rd CCL). The conference were hosted and co-organized by Shanxi University, China.

CCL is an annual conference (bi-annual before 2013) that started in 1991. It is the flagship conference of the Chinese Information Processing Society of China (CIPS), which is the largest NLP scholar and expert community in China. CCL is a premier nation-wide conference for disseminating new scholarly and technological work in computational linguistics, with a major emphasis on computational processing of the languages in China such as Mandarin, Tibetan, Mongolian, and Uyghur.

The Program Committee selected 108 papers (74 Chinese papers and 34 English papers) out of 320 submissions for publication. The acceptance rate is 33.75%. The 108 papers cover the following topics:

– Machine Translation and Multilingual Information Processing (9)
– Fundamental Theory and Methods of Computational Linguistics (7)
– Minority Language Information Processing (10)
– Social Computing and Sentiment Analysis (9)
– Text Generation and Summarization (7)
– Information Retrieval, Dialogue and Question Answering (7)
– Language Resource and Evaluation (16)
– Knowledge Graph and Information Extraction (13)
– Large Language Models (13)
– NLP Applications (17)

The final program for the 23rd CCL was the result of intense work by many dedicated colleagues. We want to thank, first of all, the authors who submitted their papers, contributing to the creation of the high-quality program. We are deeply indebted to all the Program Committee members for providing high-quality and insightful reviews under a tight schedule, and extremely grateful to the sponsors of the conference. Finally, we extend a special word of thanks to all the colleagues of the Organizing Committee and secretariat for their hard work in organizing the conference, and to the publish team, ACL Anthology and Springer for their assistance in publishing the proceedings in due time.

We thank the Program and Organizing Committees for helping to make the conference successful, and we hope all the participants enjoyed the CCL conference and a refreshing summer in Taiyuan.

July 2024

Maosong Sun, Jiye Liang, Xianpei Han, Zhiyuan Liu, Yulan He

# Organizers

## Program Committee

### 23rd CCL Conference Chairs

| | |
|---|---|
| Maosong Sun | Tsinghua University, China |
| Jiye Liang | Shanxi University, China |

### 23rd CCL Program Chairs

| | |
|---|---|
| Xianpei Han | Institute of Software , CAS, China |
| Zhiyuan Liu | Tsinghua University, China |
| Yulan He | King's College London |

### 23rd CCL Local Arrangement Chairs

| | |
|---|---|
| Hu Zhang | Shanxi University, China |
| Zhiqiang Wang | Shanxi University, China |

### 23rd CCL Evaluation Chairs

| | |
|---|---|
| Hongfei Lin | Dalian University of Technology, China |
| Hongye Tan | Shanxi University, China |
| Bin Li | Nanjing Normal University |

### 23rd CCL Publication Chairs

| | |
|---|---|
| Gaoqi Rao | Beijing Language and Culture University, China |
| Yubo Chen | Institute of Automation, CAS, China |
| Zhiliang Tian | National University of Defense Technology, China |

### 23rd CCL Chair of Frontier Forum

| | |
|---|---|
| Xin Zhao | Renmin University of China, China |

### 23rd CCL Workshop Chairs

| | |
|---|---|
| Peng Zhang | Tianjin University, China |
| Richong Zhang | Beihang University, China |

### 23rd CCL Sponsorship Chairs

| | |
|---|---|
| Xipeng Qiu | Fudan University, China |
| Binyang Li | University of International Relations, China |

**23rd CCL Publicity Chair**

| | |
|---|---|
| Jiajun Zhang | Institute of Automation, CAS, China |
| Libo Qin | Central South University, China |

**23rd CCL Website Chair**

| | |
|---|---|
| Zhenghao Liu | Northeastern University, China |

**23rd CCL System Demonstration Chairs**

| | |
|---|---|
| Peng Li | Tsinghua University, China |
| Ning Ma | Northwest Minzu University, China |

**23rd CCL Student Seminar Chairs**

| | |
|---|---|
| Tao Gui | Fudan University, China |
| Shaolei Zhang | Institution of Computing Technology, CAS, China |

**23rd CCL Tutorial Chairs**

| | |
|---|---|
| Weinan Zhang | Harbin Institute of Technology, China |
| Pengfei Liu | Shanghai Jiaotong University, China |

**23rd CCL Finance Chair**

| | |
|---|---|
| Yuxing Wang | Tsinghua University, China |

# Table of Content

*MITF：基于图像映射文本特征的跨模态图文检索方法*

娄馨月，李铀，齐睿，陈钰枫，徐金安……1

*基于 ChatGPT 查询改写的文档检索方法*

李澳，涂新辉，熊英豪……15

*基于汉语字词资源的检索增强生成与应用评估*

殷雅琦，刘扬，王悦，梁启亮……27

*面向 CQL 的语料库检索引擎的高效实现*

刘廷超，鲁鹿鸣，杨麟儿，王雨……46

*NNP-TDGM: 基于最近邻提示表征的术语 DEF 生成模型*

沈思嘉，王裴岩，王胜任，王立帮……57

*SpanCS：面向跨语言代码生成的片段级语码转换*

朱庆福，周士祺，王硕，张致铭，王昊钰，陈麒光，车万翔……71

*场景图增强的视觉语言常识推理生成*

袁凡，李丕绩……84

*基于逻辑推理和多任务融合的认知刺激对话生成方法*

蒋玉茹，李梦媛，陶宇阳，区可明，佘泽鹏，施水才……98

*基于思维链的跨语言多文档摘要生成技术研究*

祁天，杨建安， 赵铁军， 杨沐昀……110

*面向语言学习者的跨语言反馈评语生成方法*

安纪元，朱琳，杨尔弘……134

*文本样式和主题框架引导下的大模型辅助儿童新闻生成*

杜晓蒙， 于东，刘鹏远……150

*Ko-LLaMA：基于 LLaMA 的朝鲜语大语言模型*

庞杰，闫晓东，赵小兵……171

*TiComR：基于提示的藏文对话型阅读理解模型*

朋毛才让，孙媛……183

*TiLamb：基于增量预训练的藏文大语言模型*

庄文浩， 孙媛，赵小兵……197

*基于蒙古文文本语义辅助的噪声鲁棒蒙古语语音情感识别方法研究*

刘欢，梁凯麟，左昊麟，刘瑞……206

*基于神经编解码语言模型的老挝语韵律建模方法*

易宁静，王琳钦，高盛祥，余正涛……217

*基于通用依存句法的锡伯语句法树库构建研究*

周贺……230

*基于字节对编码的端到端藏语语音识别研究*

蔡郁青，王超，仁增多杰，朱宇雷，张瑾，尼玛扎西……242

*面向对话式阅读理解的高质量藏语数据集构建*

达哇才仁，朋毛才让，孙媛……254

*面向心理健康咨询的藏语数据集及大语言模型构建*

朱孟笑，沙九，冯冲……268

*融合多元特征表示的藏文命名实体识别方法*

俄见才让，周毛克，陈波，赵小兵……280

*基于对比学习和排名一致性的古代汉语翻译质量评估模型*

李怀明，邵艳秋，李炜……290

*基于两种新颖辅助任务的端到端语音翻译*

窦怀厦，吕孟哲，李军辉……305

*基于隐性句逗号识别的汉语长句机器翻译*

张文娟，李熳佳，冯文贺……314

*基于知识蒸馏的低频词翻译优化策略*

郭逸帆，咎红英，阎子悦，许鸿飞……326

*融合确定性因子及区域密度的k-最近邻机器翻译方法*

齐睿，石响宇，满志博，徐金安，陈钰枫……340

*PGA-SciRE：基于大语言模型的数据增强框架进行科学领域的关系抽取*

周洋，单世民，魏宏夔，赵哲焕，冯文铄……352

*UFSC：基于统一特征空间构建的零样本关系抽取*

刘雨辰，段建勇，孙康，张晴，何丽， 王昊，刘杰……370

*多机制整合的中文医疗命名实体识别*

王珊珊，张焜元，闫蓉……382

*基于两阶段提示学习的少样本命名实体识别*

邵佳兴，黄琪，肖聪，刘璟，罗文兵，王明文……394

*面向工艺文本的实体与关系最近邻联合抽取模型*

杨丹清忻，王裴岩，徐立军……406

*融合扩展语义和标签层次信息的文档级事件抽取*

符玉娇，廖健，李旸，郭张峰，王素格……418

*融合领域词汇扩充的低资源法律文书命名实体识别*

帕尔哈提·吐拉江，孙媛媛，蔡艾辰，王艳华，林鸿飞……431

*基于动态提示学习和依存关系的生成式结构化情感分析模型*

贾银涛，崔佳佳，穆玲玲，昝红英……442

*基于方面引导的图文渐进融合的多模态方面级情感分析方法*

闫自达，郭军军，余正涛……454

*基于联邦知识蒸馏的跨语言社交媒体事件检测*

周帅帅，朱恩昌，高盛祥，余正涛，线岩团，赵子霄，陈霖……467

*基于生成式语言模型的立场检测探究*

张袁硕，李澳华，尹召宁，王潘怡，陈波， 赵小兵……481

*基于双图注意力网络的篇章级散文情绪变化分析方法*

李爱琳，李旸，王素格，李书琪……492

*基于主题模型与图神经网络的突发公共卫生事件国际舆情演化分析研究*

高境健，桑国明，刘智，张益嘉，林鸿飞……504

*面向社交媒体多特征增强的药物不良反应检测*

李浩，邱云志，林鸿飞……515

*面向中文文本的情绪持有者抽取研究*

孙亚伟，石宇，韩旭……526

*AutoRG:一种大小模型协同的自动报告生成框架*

张京，舒江明，张宇翔，吴斌，王巍，于剑，桑基韬……539

*基于本体信息增强的人类表型概念识别*

祁杰蔚，罗凌，杨志豪，王健，林鸿飞……553

*基于机器学习的语音情感声学特征筛选*

董文琪，王涵，张璟玮……568

*基于交互行为语义模式增强的 ID 推荐方法*

王远来，白宇， 廉鹏……577

*基于双层语义映射的大语言模型辅助古汉语事件抽取半自动标注框架*

卫聪聪，李炜，冯振冰，邵艳秋……588

*基于文本风格迁移的中文性别歧视文本去毒研究*

彭健 ，左家莉，谭景璇，万剑怡，王明文……600

*基于问题扩展的散文答案候选句抽取方法研究*

雷洋，王素格，李书琪，王浩……613

*基于预训练模型与序列建模的音素分割方法*

杨尚龙，余正涛，王文君，董凌，高盛祥……625

*近三十年域外汉籍研究的现状与展望——基于文献计量分析和知识图谱绘制*

唐榕骏，彭志峰……637

*面向中文多方对话的机器阅读理解研究*

蒋玉茹，李宇，那婷婷，张仰森……650

*融合半监督学习与同义计算的传染病名称自动映射研究*

宋培彦，杨青香，胡博深，杜博雅……662

*中文语法纠错的多轮解码方法研究*

王晓盈，穆玲玲，许鸿飞……673

*中西谚语多元价值观资源库建设及对比研究*

杜霞，刘鹏远，于东……688

*从句子图到篇章图——基于抽象语义表示的篇章级共指标注体系设计*

张艺璇，李斌，许智星，卢芃秀……700

*汉语中介语词同现网络研究*

钱隆，赵慧周，丁芊，王治敏……712

*基于意合图语义理论的结构标注体系与资源建设*

郭梦溪，李梦，荀恩东，饶高琦，于钟洋……726

*意合图：中文多层次语义表示方法*

郭梦溪，荀恩东，李梦，饶高琦……740

*由 L2 到 L1 的跨语言激活路径研究——基于词汇识别的 ERP 数据*

杨思琴，胡美，江铭虎……750

*大语言模型故事理解能力评价数据集*

闫国航，郭亚鑫，谭红叶，张虎……760

*大语言模型开放性生成文本中的职业性别偏见研究*

张旭，郭梦清，朱述承，于东，刘颖，刘鹏远……774

*大语言模型在中文文本纠错任务的评测*

穆玲玲，王晓盈，崔佳佳……790

*面向“以 A 为 B”构式语义场景的汉语框架识别数据集构建*

杨沛渊，苏雪峰，李俊材，闫智超，柴清华，李茹……807

*上古汉语分词和词性标注语料库的构建*

柯永红……819

*图解句式结构体系及其树库构建*

彭炜明，赵敏，宋昱辰，胡佳佳，宋天宝，穗志方，宋继华……830

*英语科技论文摘要语步结构语料库构建研究*

李洪政，王若锦，冯冲，刘芳……841

*Self-Guide:一种基于自我规划的大语言模型推理增强方法*

刘艺彬，刘正皓，闫宇坤，于是，王硕，杨麟儿，陈慧敏，谷峪，于戈……853

*基于大模型的交互式谎言识别：数据和模型*

纪程炜，王思远，李太山，牟馨忆，赵丽敏，薛兰青，应缜哲，王维强，黄萱菁，魏忠钰……870

*基于动态聚类与标签空间映射的上下文学习模板构建方法*

张琦，金醒男，裴誉，杜永萍……883

*基于领域信息分解式学习的大语言模型修辞认知增强方法*

王雯，于东，刘鹏远……894

*基于中间层对齐的异构师生模型知识蒸馏*

翟飞燕，王任之，李丕绩……910

*面向小规模大语言模型推理优化的推理路径排序方法*

李俊，白宇，刘雨婷……929

*面向中文实体识别的 Transformers 模型句子级非对抗鲁棒性研究*

王立帮，王裴岩，沈思嘉……941

*银瞳:基于自适应语义空间学习的中文金融多任务大模型*

周宇航，李泽平，田思雨，倪雨琛，张健，刘响，叶广楠，吴杰，柴洪峰……955

*Enhancing Free-Form Table Question Answering Models by Distilling Relevant-Cell-Based Rationales*

Zhiyu Yang, Shuo Wang, Yukun Yan, Pengyuan Liu, Dong Yu……973

*Enhangcing Sequence Representation for Personalized Search*

Shijun Wang, Han Zhang, Zhe Yuan……986

*Joint Similarity Guidance Hash Coding Based on Adaptive Weight Mixing Strategy For Cross-Modal Retrieval*

Yaqi Sun, Jing Yun, Zhuoqun Ma…… 999

*Generate-then-Revise: An Effective Synthetic Training Data Generation framework for Event Detection*

Huidong Du, Hao Sun, Pengyuan Liu, Dong Yu……1011

*E³: Optimizing Language Model Training for Translation via Enhancing Efficiency and Effectiveness*

Linqing Chen, Weilei Wang, and Dongyang Hu……1023

*Multi-features Enhanced Multi-task Learning for Vietnamese Treebank Conversion*
Zhenguo Zhang, Jianjian Liu, Ying Li····· 1035
*SimCLNMT: A Simple Contrastive Learning Method for Enhancing Neural Machine Translation Quality*
Menglong Xu, Yanliang Zhang·····1047
*Translate-and-Revise: Boosting Large Language Models for Constrained Translation*
Pengcheng Huang, Yongyu Mu, Yuzhang Wu, Bei Li, Chunyang Xiao, Tong Xiao, and Jingbo Zhu·····1059
*A Multi-Task Biomedical Named Entity Recognition Method Based on Data Augmentation*
Hui Zhao, Di Zhao,Jiana Meng, Shuang Liu, Hongfei Lin·····1075
*Biomedical Event Causal Relation Extraction by Reasoning Optimal Entity Relation Path*
Lishuang Li, Liteng Mi, Beibei Zhang, Yi Xiang, Yubo Feng, Xueyang Qin, Jingyao Tang·····1087
*Joint Entity and Relation Extraction Based on Bidirectional Update and Long-Term Memory Gate Mechanism*
Yili Qian，Enlong Ren，Haonan Xu·····1099
*MFE-NER:Multi-feature Fusion Embedding for Chinese Named Entity Recognition*
Jiatong Li, Kui Meng·····1112
*UDAA: An Unsupervised Domain Adaptation Adversarial Learning Framework for Zero-Resource Cross-Domain Named Entity Recognition*
Baofeng Li,Jianguo Tang,Yu Qin,Yuelou Xu, Yan Lu, Kai Wang, Lei Li, Yanquan Zhou·····1123
*Triple-view Event Hierarchy Model for Biomedical Event Representation*
Jiayi Huang, Lishuang Li, Xueyang Qin, Yi Xiang, Jiaqi Li, Yubo Feng·····1136
*DialectMoE: An End-to-End Multi-Dialect Speech Recognition Model with Mixture-of-Experts*
Jie Zhou，Shengxiang Gao，Zhengtao Yu, Ling Dong, Wenjun Wang·····1148
*Distinguishing Neural Speech Synthesis Models Through Fingerprints in Speech Waveforms*
Chu Yuan Zhang, Jiangyan Yi, Jianhua Tao, Chenglong Wang, Xinrui Yan·····1160
*Knowledge Graph-Enhanced Recommendation with Box Embeddings*
Qiuyu Liang, Weihua Wang, Lei Lv, Feilong Bao·····1172
*Readability-guided Idiom-aware Sentence Simplification(RISS) for Chinese*
Jingshen Zhang, Xinglu Chen, Xinying Qiu, Zhimin Wang, Wenhe Feng·····1183
*A Tone-based Hierarchical Structure of Chinese Prosody*
Ya Li····· ···1201
*Linguistic Guidance for Sequence-to-Sequence AMR Parsing*
Binghao Tang,Boda Lin,Si Li·····1212
*Automatic Construction of the English Sentence Pattern Structure Treebank for Chinese ESL learners*
Lin Zhu, Meng Xu, Wenya Guo, Jingsi Yu, Liner Yang, Zehuang Cao, Yuan Huang, Erhong Yang·····1223
*Cost-efficient Crowdsourcing for Span-based Sequence Labeling:Worker Selection and Data Augmentation*

Yujie Wang, Chao Huang, Liner Yang, Zhixuan Fang, Yaping Huang, Yang Liu, Jingsi Yu, Erhong Yang……1239

*DLUE: Benchmarking Document Language Understanding*

Ruoxi Xu, Hongyu Lin, Xinyan Guan, Yingfei Sun, Le Sun……1257

*Do Large Language Models Understand Conversational Implicature- A case study with a Chinese sitcom*

Shisen Yue, Siyuan Song, Xinyuan Cheng, Hai Hu……1270

*EmoFake: An Initial Dataset for Emotion Fake Audio Detection*

Yan Zhao, Jiangyan Yi, Jianhua Tao, Chenglong Wang, Yongfeng Dong……1286

*Going Beyond Passages:Readability Assessment for Book-level Long Texts*

Wenbiao Li, Rui Sun, Tianyi Zhang, Yunfang Wu……1298

*Mitigating the Bias of Large Language Model Evaluation*

Hongli Zhou, Hui Huang, Yunfei Long, Bing Xu, Conghui Zhu, Hailong Cao, Muyun Yang, Tiejun Zhao……1310

*PPDAC: A Plug-and -Play Data Augmentation Component for Few-shot Extractive Question Answering*

Qi Huang, Han Fu, Wenbin Luo, Mingwen Wang, Kaiwei Luo……1320

*Sentence-Space Metrics (SSM) for the Evaluation of Sentence Comprehension*

Jieyu Lin, Honghua Chen, Nai Ding……1334

*AuditWen: An Open-Source Large Language Model for Audit*

Jiajia Huang, Haoran Zhu, Chao Xu, Tianming Zhan, Qianqian Xie, Jimin Huang……1351

*Chinese Grammatical Error Correction via Large Language Model Guided Optimization Training*

Xiao Liu, Ying Li, Zhengtao Yu……1366

*Pattern Shifting or Knowledge Losing? A Forgetting Perspective for Understanding the Effect of Instruction Fine-Tuning*

Chunkang Zhang, Boxi Cao, Yaojie Lu, Hongyu Lin, Liu Cao, Ke Zeng, Guanglu Wan, Xunliang Cai, Xianpei Han, Le Sun……1381

*Prior Constrains-based Reward Model Training for Aligning Large Language Models*

Hang Zhou, Chenglong Wang, Yimin Hu, Tong Xiao, Chunliang Zhang, and Jingbo Zhu…… 1395

*Prompt Engineering 101 Prompt Engineering Guidelines from a Linguistic Perspective*

Wenjuan Han, Xiang Wei, Xingyu Cui, Ning Cheng, Guangyuan Jiang, Weinan Qian, Chi Zhang……1408

# MITF：基于图像映射文本特征的跨模态图文检索方法

娄馨月[1,2]，李铀[1,2]，齐睿[1,2]，陈钰枫[1,2]，徐金安[1,2*]
[1]北京交通大学，计算机科学与技术学院，北京，100044
[2]交通数据分析与挖掘北京市重点实验室，北京，100044
{20241254,21241143,20281284,chenyf,jaxu}@bjtu.edu.cn

## 摘要

减小图文信息间的语义鸿沟，促进跨模态信息的对齐与融合一直是解决跨模态图文检索问题的关键。但现有的双流模型因为训练时图像编码器与文本编码器是分开的，导致图文特征的对齐与融合较难。因此，本文提出图像映射文本特征（MITF）网络将不同模态（图像和文本）的信息映射到单一模态（文本），进一步增强跨模态语义的融合和对齐，提高图文检索的性能。具体地，在冻结预训练的中文视觉语言模型Chinese-CLIP参数的情况下，训练一个MITF网络将图像映射为伪语言标记，在此基础上引入提示词自动学习机制提升模型对于伪语言标记的理解能力。同时，在检索时构建Faiss索引提高检索速度。在三个开源数据集的实验结果表明所提方法相比原始Chinese-CLIP模型检索时的Mean Recall指标平均提高了3.7%，检索速度提高了约4倍。同时，图文特征可视化结果进一步表明所提方法提高了图像特征与文本特征的对齐程度。

**关键词：** 语义对齐；跨模态；图文检索；图像映射文本特征网络

# MITF:Cross-modal Image-text Retrieval Method with Mapping Images to Text Features

**Xinyue Lou[1,2], You Li[1,2], Rui Qi[1,2], Yufeng Chen[1,2], Jinan Xu[1,2*]**
[1]School of Computer Science and Technology, Beijing Jiaotong University, Beijing, 100044
[2]Beijing Key Lab of Traffic Data Analysis and Mining, Beijing, 100044
{20241254,21241143,20281284,chenyf,jaxu}@bjtu.edu.cn

## Abstract

To address the cross-modal image-text retrieval challenge, it's essential to close the semantic gap between vision and language. However, the prevailing dual-stream model which separates image encoder and text encoder during training complicates the problem. In this paper, we propose a mapping network to map information from different modalities (image and text) to a single modality (text), called MITF network, with the purpose of improving the fusion and alignment of cross-modal semantics. Given a frozen pre-trained model, Chinese-CLIP, we train the MITF network to convert the visual embedding into the corresponding pseudo language tokens. Additionally, we also introduce a mechanism that automatically learns prompt to enhance the model's understanding of pseudo language tokens. To speed up the retrieval process, the Faiss index is employed. Experimental findings on three datasets show that our method achieves an absolute 3.7% MR gain and quadruples retrieval speed compared to the

*徐金安（通讯作者）：jaxu@bjtu.edu.cn

第二十三届中国计算语言学大会论文集，第1页-第14页，太原，中国，2024年7月25日至28日。
卷1：主会论文
(c) 2024 中国中文信息学会计算语言学专业委员会

vanilla Chinese-CLIP model. Additionally, the image-text feature visualization results further show that our method improves the alignment between image features and text features.

**Keywords:** Semantic alignment , Cross-modal , Image-text retrieval , Mapping images to text features network

## 1 引言

近年来，随着图像、文本等多模态数据的高速增长，对于访问和利用各种跨模态数据的需求也相应增长，形成了跨模态研究的三个主要领域：跨模态生成、跨模态视觉问答和跨模态检索。跨模态检索是指给定一种模态样本，返回另一模态中相似的样本，如图像-文本检索等，在公安、医疗和电商领域等均有着广泛的应用前景(Kaur et al., 2021)。但由于各模态语义间存在的差异，如何减小差异实现跨模态语义的有效对齐逐渐成为了研究热点。

受预训练语言模型成功的启发(Devlin et al., 2018; Liu et al., 2019; Brown et al., 2020)，大规模预训练视觉语言模型研究(Li et al., 2019a; Radford et al., 2021; Wang et al., 2022a; Liu et al., 2023a)在各种跨模态任务上取得了显著进展，包括图像-文本检索。这些现有方法通常可以根据模型架构大致分为两类：单流模型和双流模型。单流模型通常会添加额外的Transformer层(Vaswani et al., 2017)来模拟图像和文本表示之间的深度交互，从而提高检索性能，如图1(a)所示。但当该模型应用于整个图像集合时会导致检索速度不稳定，因为每当给出新的文本查询时，每个图像样本都需要进行跨模态计算。相比之下，双流模型以完全解耦的方式编码视觉和文本输入，如图1(b)所示。图像表示允许预先计算并独立于文本查询重新使用，检索速度较快。但双流模型也因为图像编码器和文本编码器在训练时参数是独立的，导致两个编码器对图文信息编码后得到的特征不是完全对齐的，多模态语义信息的融合和交互较难。虽然存在一些权衡方法，如ALBEF模型(Li et al., 2021)通过跨模态注意力机制，将图像表征和文本表征先对齐再进行融合，但训练时计算成本较高，且检索速度较慢。

为了解决以上挑战，本文围绕跨模态检索中的语义对齐问题，提出图像映射文本特征（Mapping Images to Text Features，MITF）网络，如图1(c)所示，将图像映射为文本特征后，在文本编码器端模拟跨模态信息间的深度交互，进一步增强跨模态语义的有效对齐和融合，从而提高检索性能。同时，图像表示允许预先计算，检索速度较为稳定。具体来说，首先在冻结预训练的中文视觉语言模型Chinese-CLIP（以下简称CN-CLIP）模型(Yang et al., 2022)参数的情况下，训练一个MITF网络将图像映射为伪语言标记即用文本特征表示图像信息的关键内容。而为了使伪语言标记更好地为文本编码器所利用，引入提示词（prompt）自动学习机制进行优化。然后在检索时构建Faiss索引，有效提升检索速度。该方法的有效性在Flickr30k-CN(Lan et al., 2017)数据集、COCO-CN(Li et al., 2019b)数据集和MUGE(Lin et al., 2021)数据集上均得到了验证。本文的主要贡献总结如下：

- 提出图像映射文本特征网络，并引入提示词自动学习机制，利用预训练的视觉语言模型将图像转换为伪语言标记，以减小图文信息间的语义鸿沟。
- 在检索时构建Faiss索引，提高检索效率，在检索精度不变的情况下，检索速度提高了约四倍。
- 经过实验验证，本文所提出的图像映射文本特征网络以较小的参数量实现了与全量微调方法相当的结果，有效降低了模型训练的计算资源和成本。

## 2 相关工作

### 2.1 跨模态图文检索

近年来，随着注意力机制的广泛应用，预训练视觉语言模型已被广泛研究并应用于图文检索领域(Tan and Bansal, 2019; Wang et al., 2022a; Liu et al., 2023a)。模型结构大致可分

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

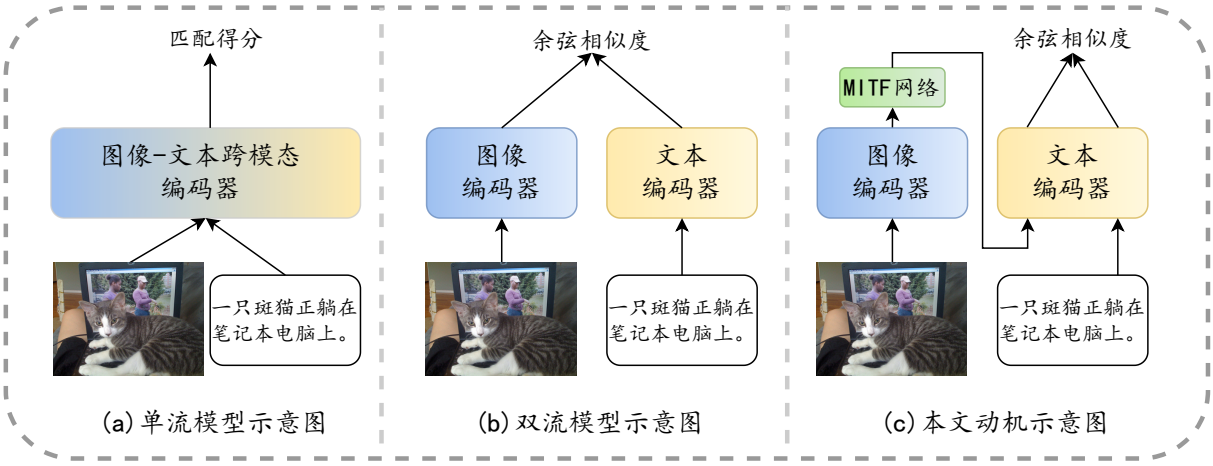



图 1: 图文检索模型与本文动机示意图

为两类：单流模型和双流模型。单流模型通过深度交互模块对图像和文本进行联合编码，并输出融合特征。Lu et al. (2019)提出的早期算法采用对象检测器提取图像特征，但通常忽略了重要的背景信息。随后，Kim et al. (2021)提出ViLT模型将图像编码器和文本编码器统一为Transformer模型(Vaswani et al., 2017)，以充分利用所有信息。然而，这些模型都依赖于跨模态Transfomer编码器来同时融合视觉和文本信号，需要大量的计算成本并降低了推理速度。双流模型主要侧重于学习如何将从独立编码器中获得的视觉特征与文本特征相统一。由于只有一个轻量级交互模块（通常是MLP网络或点乘法）用于匹配图像和文本特征，因此双流模型可以在数十亿实例上进行对比学习，如CLIP模型(Radford et al., 2021)、ALIGN模型(Jia et al., 2021)和BLIP2模型(Li et al., 2023)等。在双流模型中，不同模态的交互只发生在最终的轻量级模块，导致与单流模型相比性能稍差。然而，这种双流模型的后期交互方案允许图像和文本表示的预先计算，从而使它适合于实时搜索。

通常，双流模型由于缺乏有效的跨模态特征融合，表现出较差的检索性能。而单流模型通过深度交互模块在编码过程中整合来自多种模态的信息，通常会导致检索性能较好，但牺牲了灵活性导致检索速度极低。此后虽然一些权衡方法被提出以解决以上问题，如LiT方法(Zhai et al., 2022)训练时冻结图像编码器以整合较为干净的图像数据源和图文对数据源的优点，而ALBEF模型(Li et al., 2021)则通过跨模态注意力机制，先将图像表征和文本表征对齐再融合，但训练时的计算成本较高。受ALBEF模型的启发，本文在CN-CLIP模型(Yang et al., 2022)的基础上，提出图像映射文本特征网络，试图将多模态信息映射到文本模态以减小模态间的差异，进一步增强跨模态语义的对齐和融合。

### 2.2 用语言标记表示图像方法

曾经有一些研究试图在视觉语言模型的预训练过程中将图像区域映射为语言标记(Wang et al., 2022b; Chen et al., 2019; Li et al., 2020)。典型的框架包括：(1) 使用预训练的物体检测器检测图像中的物体；(2) 将图像区域和相应的文本描述输入文本编码器；(3) 优化多个多模态目标以获得视觉语言模型。但这些方法都需要在预训练阶段使用高性能的物体检测器。还有其他方法如Cohen et al. (2022)使用循环对比损失来学习如何将一组图像转换成一个概念词标记，以解决个性化图像检索和语义分割等问题。Gal et al. (2022)提出用单词标记来表示描绘同一对象的几幅图像，并将其用于文本到图像的生成任务。Saito et al. (2023)提出pic2word方法，使用未标记图像数据集进行无监督对比学习，以便灵活地组合图像和文本查询。受这些方法的启发，本文提出将图像映射为伪语言标记的方法，以实现跨模态语义更有效的对齐，提升图文检索的效率。

## 3 方法

在本节中，将介绍本文的MITF网络和提示词自动学习机制及检索方法。训练和推理过程的概述如图2所示。所采用的基座模型为预训练的视觉语言模型CN-CLIP(Yang et al.,


(c) 2024 中国中文信息学会计算语言学专业委员会

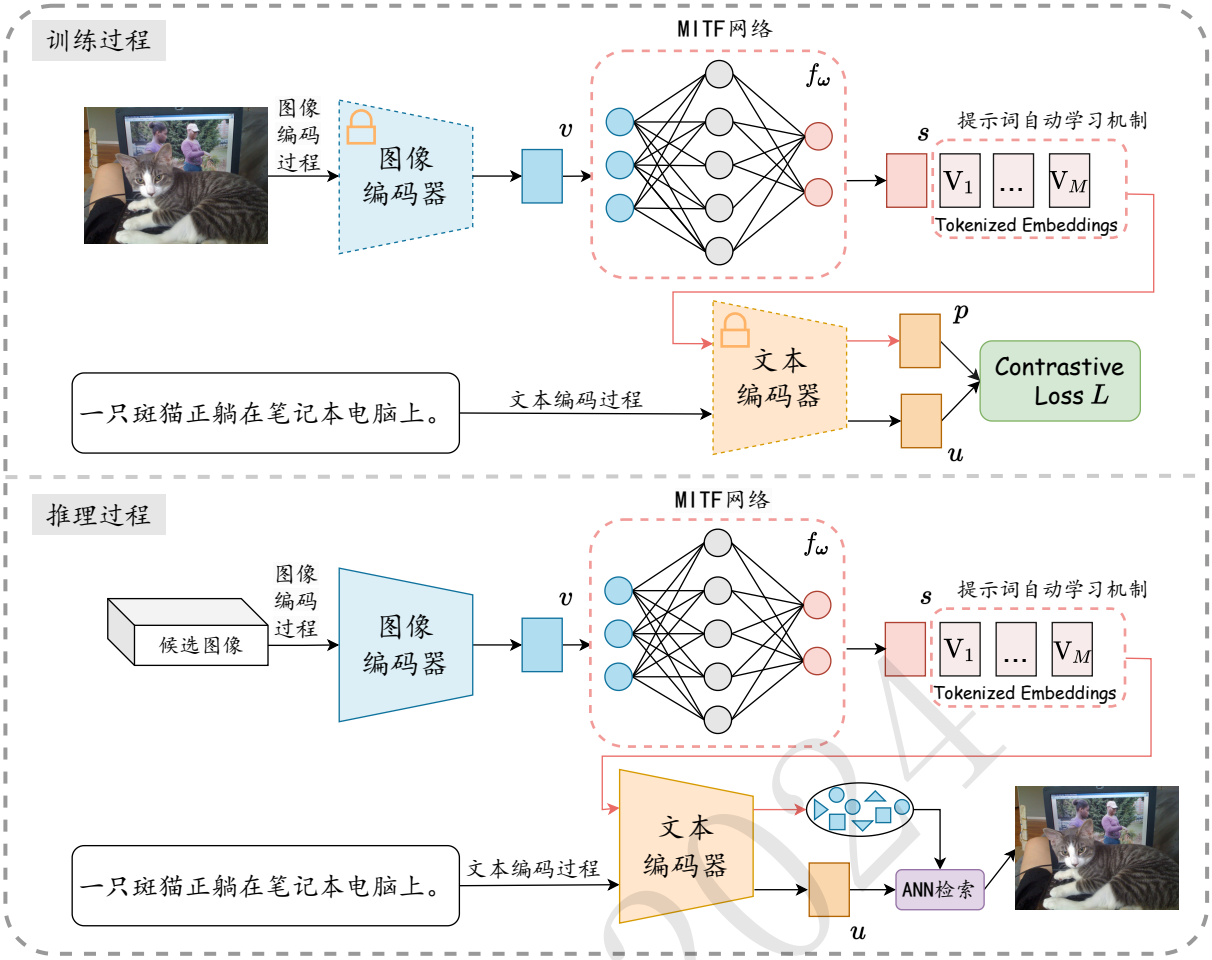



图 2: 训练过程和推理过程概述图

2022)，该模型由文本编码器和图像编码器组成，一共提供了五个不同规模的模型。文本编码器为RoBERTa-wwm系列模型(Liu et al., 2019)，图像编码器为RN50模型(He et al., 2016)或者Vision Transformer系列模型(Dosovitskiy et al., 2020)。在训练时，冻结预训练文本编码器和图像编码器，训练一个映射网络将视觉嵌入转换为相应的伪语言标记。并引入提示词自动学习机制使得到的伪语言标记更好地为文本编码器所利用。在推理时，将图像经过MITF网络得到伪语言标记，将其与待检索文本分别经过文本编码器得到的特征进行比较，通过近似近邻检索得到最相似的图文样本对。我们将在以下小节中详细介绍训练和推理机制。

### 3.1 对比语言-图像预训练模型（CN-CLIP）

给定一个图像文本数据集$\mathcal{D} = \{(\boldsymbol{x}_n, \boldsymbol{t}_n)\}_{n=1}^{N}$，其中$x \in \mathcal{X}$为图像，$t \in \mathcal{T}$为分词后的语言描述。对于一张图像$x$，以$\theta$为参数的图像编码器$f_{\boldsymbol{\theta}}$对其进行编码得到视觉表征$\tilde{\boldsymbol{v}} \in R^{d\times 1}$：$\tilde{\boldsymbol{v}} = f_{\boldsymbol{\theta}}(\boldsymbol{x})$。对于文本描述$t$，以$\phi$为参数的文本编码器$f_{\boldsymbol{\phi}}$对其进行编码提取语言表征$\tilde{\boldsymbol{u}} \in R^{d\times 1}$：$\tilde{\boldsymbol{u}} = f_{\boldsymbol{\phi}}(\boldsymbol{t})$。CN-CLIP模型(Yang et al., 2022)采用对比学习的方法，使用约2亿对中文图文对进行训练。对于批次$\mathcal{B}$中的第$i$个图像$\boldsymbol{x}_i$和与其对应的第$j$个文本描述$\boldsymbol{t}_j$，对其分别进行归一化处理$\boldsymbol{v}_i = \frac{\tilde{\boldsymbol{v}}_i}{\|\tilde{\boldsymbol{v}}_i\|}$和$\boldsymbol{u}_j = \frac{\tilde{\boldsymbol{u}}_j}{\|\tilde{\boldsymbol{u}}_j\|}$。最后，CN-CLIP模型使用对称交叉熵损失进行优化，损失函数如下所示：

$$\min_{\{\boldsymbol{\theta},\boldsymbol{\phi}\}} \mathcal{L}_{\text{con}} = \mathcal{L}_{t2i} + \mathcal{L}_{i2t} \tag{1}$$


(c) 2024 中国中文信息学会计算语言学专业委员会

其中包括2个对比项，温度系数$\tau$用于控制对困难负样本的惩罚力度，如下式所示：

$$\mathcal{L}_{t2i} = -\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\log\frac{\exp\left(\tau \boldsymbol{u}_i^T \boldsymbol{v}_i\right)}{\sum_{j\in\mathcal{B}}\exp\left(\tau \boldsymbol{u}_i^T \boldsymbol{v}_j\right)} \tag{2}$$

$$\mathcal{L}_{i2t} = -\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\log\frac{\exp\left(\tau \boldsymbol{v}_i^T \boldsymbol{u}_i\right)}{\sum_{j\in\mathcal{B}}\exp\left(\tau \boldsymbol{v}_i^T \boldsymbol{u}_j\right)} \tag{3}$$

而在进行图文检索推理时，用余弦相似度来衡量图像文本的相似度大小，给定一张图像$x$，预测文本描述$t$与其匹配的概率公式为：

$$p(t=i \mid \boldsymbol{x}) = \frac{\exp\left(\cos\left(\boldsymbol{u_i}, \boldsymbol{v}\right)/\tau\right)}{\sum_{j=1}^{N}\exp\left(\cos\left(\boldsymbol{u_j}, \boldsymbol{v}\right)/\tau\right)} \tag{4}$$

同理，给定一个文本描述$t$，预测图像$x$与其匹配的概率公式为：

$$p(x=i \mid \boldsymbol{t}) = \frac{\exp\left(\cos\left(\boldsymbol{v_i}, \boldsymbol{u}\right)/\tau\right)}{\sum_{j=1}^{N}\exp\left(\cos\left(\boldsymbol{v_j}, \boldsymbol{u}\right)/\tau\right)} \tag{5}$$

### 3.2 图像映射文本特征（MITF）网络

该方法的核心是训练一个映射网络来输出伪语言标记，在固定文本编码器和图像编码器的情况下训练该网络，这样获得的伪语言标记能够最大程度地减少图像特征和文本特征之间的对比损失，即强制网络形成一个从图像嵌入到语言嵌入的循环，如图2所示，将图文信息都映射到文本模态以减小模态间特征的差异。对于未归一化的视觉嵌入$\tilde{\boldsymbol{v}}$，参数为$\boldsymbol{\omega}$的映射网络$f_{\boldsymbol{\omega}}$提取伪语言标记嵌入$s = f_{\boldsymbol{\omega}}(\tilde{\boldsymbol{v}})$。$f_{\boldsymbol{\omega}}$为一个深度为3层的多层感知机，其参数大约为17M。如图2右侧所示，在$s$后添加自动优化的提示句（关于提示句优化的设计在3.3章节中介绍），得到$\hat{s}$。然后将$\hat{s}$输入文本编码器$f_{\boldsymbol{\phi}}$，得到伪语言嵌入$\tilde{\boldsymbol{p}} = f_{\phi}(\hat{s})$，希望$\tilde{\boldsymbol{p}}$能够在表示输入的图像信息的基础上尽可能接近于其对应的原文本的语言嵌入$\tilde{\boldsymbol{u}}$。该映射网络的对比损失为：

$$\min_{M}\mathcal{L} = \mathcal{L}_{t2i}(\boldsymbol{p}, \boldsymbol{u}) + \mathcal{L}_{i2t}(\boldsymbol{p}, \boldsymbol{u}) \tag{6}$$

其中包括2个对比项，如下所示：

$$\mathcal{L}_{t2i}(\boldsymbol{p}, \boldsymbol{u}) = -\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\log\frac{\exp\left(\tau \boldsymbol{u}_i^T \boldsymbol{p}_i\right)}{\sum_{j\in\mathcal{B}}\exp\left(\tau \boldsymbol{u}_i^T \boldsymbol{p}_j\right)} \tag{7}$$

$$\mathcal{L}_{i2t}(\boldsymbol{p}, \boldsymbol{u}) = -\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\log\frac{\exp\left(\tau \boldsymbol{p}_i^T \boldsymbol{u}_i\right)}{\sum_{j\in\mathcal{B}}\exp\left(\tau \boldsymbol{p}_i^T \boldsymbol{u}_j\right)} \tag{8}$$

其中$\boldsymbol{p}_i = \frac{\tilde{\boldsymbol{p}}_i}{\|\tilde{\boldsymbol{p}}_i\|}$。在此基础上训练$f_{\boldsymbol{\omega}}$生成$\boldsymbol{s}_i$使$\boldsymbol{p}_i$接近$\boldsymbol{u}_i$。上述公式与式（1、2、3）的不同之处在于：（1）优化是针对参数$\boldsymbol{\omega}$进行的，而其他部分是固定的；（2）使用$p$代替$v$，换言之，对比学习的目标是使该图像生成的伪语言标记与其真实文本标记尽可能接近。

### 3.3 提示词自动学习机制

在3.2节中通过MITF网络得到的伪语言标记是连续的特征，所表达的语义信息较为受限，为了使生成的伪语言标记更好地为文本编码器所利用，并且受提示学习在预训练视觉语言模型上同样取得了令人瞩目的性能的启发(Zhou et al., 2022)，我们引入了提示词自动学习机制。提示工程通过给定具有指示性的离散字符辅助视觉语言模型完成相应的下游任务，如clip模型在进行图像分类时给定一个提示语句：*A photo of a object*（其中object为待分类的物体名称）(Radford et al., 2021)，使其完成图像分类的任务，而提示学习的研究旨在将这一过程自动化，而不是手动设计提示。提示学习的方法为我们如何更好地提升模型对于伪语言标记的理解能力以提高检索性能提供了可供借鉴的思路。它通过从数据中学习端到端连续向量对上下文词


(c) 2024 中国中文信息学会计算语言学专业委员会

语进行建模，避免手动提示调优。概述如图2所示。具体来说，提供给文本编码器$f_\phi$的提示设计为如下形式：

$$\hat{s} = [\mathrm{s}][\mathrm{V}]_1[\mathrm{V}]_2 \dots [\mathrm{V}]_M \tag{9}$$

其中，每个$[\mathrm{V}]_m (m \in \{1, \dots, M\})$是一个维度与词嵌入相同的向量（如CN-CLIP$_{\text{ViT-B}}$模型中的768维），而$\mathcal{M}$是一个超参数，指定了上下文标记的数量，$s$为MITF网络中所学习到的伪语言标记。给定一张图像$x$，预测文本描述$t$与其匹配的概率公式为：

$$p(t = i \mid \boldsymbol{x}) = \frac{\exp\left(\cos\left(\boldsymbol{u_i}, \boldsymbol{p}\right)/\tau\right)}{\sum_{j=1}^{N} \exp\left(\cos\left(\boldsymbol{u_j}, \boldsymbol{p}\right)/\tau\right)} \tag{10}$$

同理，给定一个文本描述$t$，预测图像$x$与其匹配的概率公式为：

$$p(x = i \mid \boldsymbol{t}) = \frac{\exp\left(\cos\left(\boldsymbol{p_i}, \boldsymbol{u}\right)/\tau\right)}{\sum_{j=1}^{N} \exp\left(\cos\left(\boldsymbol{p_j}, \boldsymbol{u}\right)/\tau\right)} \tag{11}$$

在交叉熵损失函数的基础上，训练过程中将进行梯度调整以最小化对称多类损失，梯度可以通过文本编码器$f_\phi$反向传播，利用参数中编码的丰富知识来优化提示词获得更好的表示。

### 3.4 使用Faiss索引加速的推理过程

在给定文本检索图像时推理过程概述如图2所示，候选图像经过图像编码器和MITF网络得到伪语言标记，将其与给定文本分别经过文本编码器得到的特征进行对比，通过检索得到最相似的k个样本对，采用余弦相似度函数来衡量样本相似度大小。对于检索过程，本文采用了Faiss框架提供的近似近邻（Approximate Nearest Neighbor，ANN）检索方法来提高检索速度。Faiss*全称是Facebook AI Similarity Search，是由Facebook开发的用于高效相似性搜索的开源库，是目前最为成熟的ANN检索库。对于检索任务，可大致分为以下三步：创建索引列表、聚类训练、查询。我们使用了Faiss的倒排乘积量化方法来加快检索速度，其思路是将原始的高维向量空间划分为多个低维子空间，在每个子空间内进行向量量化编码。最终将所有子空间量化编码依次串联起来作为原向量的编码。在查询时，需要对查询向量进行同样的子空间拆分和量化，然后基于这些量化码在倒排索引中进行近邻搜索。经过处理后，我们得到了经过乘积量化器优化和k-means训练的Faiss索引表$I$。在给定文本检索图像时的算法伪代码如下所示。对于输入向量$u$，Faiss 能快速检索出与$u$最相似的前k个备选向量，表示为$[h_1, h_2, \dots, h_k]$，即候选图像集中与给定文本最匹配的k个样本。同理可得给定图像检索文本的过程。

**Algorithm 1: 基于Faiss索引的图文检索算法伪代码**

**Input**：text_features，image_features
**Output**：top_k_image_ids
**for** image_feature in image_features **do**
    image_feats=list(image_feature)
index=faiss.IndexFlatIP()
index.add(image_feats)
**for** text_feature in text_features **do**
    indices = index.search(text_feature, top$_k$)
    **Return** indices as top_k_image_ids

## 4 实验

### 4.1 实验设置

本实验所采用的基座模型为Aliyun开源的CN-CLIP模型†，若无特别指明，模型所采用的基座模型均为CN-CLIP$_{\text{ViT-B}}$模型。MITF网络由深度为3层的多层感知机网络组成，所采用的激活函数为Mish函数(Misra, 2019)（输出层没有使用激活函数，具体的设计细节将在附录A中

*https://github.com/facebookresearch/Faiss
†https://github.com/OFA-Sys/Chinese-CLIP


(c) 2024 中国中文信息学会计算语言学专业委员会

给出）。使用的优化器为AdamW(Loshchilov and Hutter, 2017)，学习率为$8 * 10^{-4}$，权重衰减为0.1。对比学习的批量大小为576，设置迭代次数为10，采用学习率余弦衰减策略，具体的超参数设置将在附录B中给出。基座模型为CN-CLIP$_{\text{ViT-L}}$和CN-CLIP$_{\text{ViT-H}}$的模型在3块GeForce RTX 3090 GPU上进行训练，其余模型均在6块GeForce RTX 2080 Ti GPU上进行训练。本次实验所选用的数据集为Flickr30k-CN数据集(Lan et al., 2017)、COCO-CN数据集(Li et al., 2019b)和MUGE数据集(Lin et al., 2021)，关于各数据集详细数据将在附录C中给出。首先在对应数据训练集上冻结预训练模型训练MITF网络，选取在验证集上效果最好的模型，然后在对应测试集上测试结果。由于MUGE数据集官方没有提供图到文标注数据和测试集数据，关于MUGE数据集只在验证集上进行文到图检索实验。同时，我们也在相同实验环境下进行了CN-CLIP全量微调实验以进行对比，对MITF网络进行全面分析。

### 4.2 实验结果

对于图像检索文本和文本检索图像任务，本文使用该任务常用的Recall@1 (R@1)、Recall@5 (R@5)、Recall@10 (R@10) 和Mean Recall (MR)指标来进行评测。

| | 方法 | 训练参数量 | 检索时间 | 图像检索文本 | | | | 文本检索图像 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | R@1 | R@5 | R@10 | MR | R@1 | R@5 | R@10 | MR |
| COCO-CN | CN-CLIP | - | 4s | 57.0 | 84.1 | 93.6 | 78.2 | 62.2 | 86.6 | 94.9 | 81.2 |
| | Finetune* | 188M | 4s | **67.1** | **93.1** | **97.6** | **85.9** | **68.3** | **93.4** | **97.6** | **86.5** |
| | **Ours** | **17M** | **1s** | 63.7 | 92.4 | 97.2 | 84.4 | 66.9 | 91.7 | 96.8 | 85.1 |
| Flickr30k-CN | CN-CLIP | - | 9s | 74.6 | 93.5 | 97.1 | 88.4 | 62.7 | 86.9 | 92.8 | 80.8 |
| | Finetune* | 188M | 9s | **87.5** | **97.8** | **99.4** | **94.9** | **71.4** | **92.2** | **96.1** | **86.6** |
| | **Ours** | **17M** | **3s** | 84.9 | 96.9 | 99.1 | 93.6 | 70.0 | 90.9 | 95.6 | 85.5 |
| MUGE | CN-CLIP | - | 139s | - | - | - | - | 52.1 | 76.7 | 84.4 | 71.1 |
| | Finetune* | 188M | 139s | - | - | - | - | **54.2** | **80.3** | **86.9** | **73.8** |
| | **Ours** | **17M** | **41s** | - | - | - | - | 52.3 | 78.2 | 86.0 | 72.2 |
| 平均 | CN-CLIP | - | 51s | 65.8 | 88.8 | 95.4 | 83.3 | 59.0 | 83.4 | 90.7 | 77.7 |
| | Finetune* | 188M | 51s | **77.3** | **95.5** | **98.5** | **90.4** | **64.6** | **88.6** | **93.5** | **82.3** |
| | **Ours** | **17M** | **15s** | 74.3 | 94.7 | 98.2 | 89.0 | 63.1 | 86.9 | 92.8 | 80.9 |

表 1: 与CN-CLIP全量微调方法的对比（*为我们复现的实验结果），检索时间为在相应数据集的测试集（验证集）检索所有数据所需的时间

**与CN-CLIP全量微调方法的对比。**对比学习的效果与学习时的批量大小有较大关系，在内存资源一定的情况下，模型参数量越多，对比学习的批量数就越小。因此，为保证实验的公平性，我们在相同实验环境条件下对CN-CLIP$_{\text{ViT-B}}$模型进行了全量微调方法的实验，对比结果如表1所示，在3个数据集上的平均实验结果表明，训练时MITF网络参数量为全量微调参数量的9.4%，但与全量微调实验结果MR值的误差仅为1.4%。而与原始CN-CLIP模型零样本预测的表现相比，本文所提出的模型在3个数据集上均达到了较好的效果，在COCO-CN数据集上相比原始CN-CLIP模型MR指标提升了5.0%，在Flickr30k-CN数据集上提升了5.0%，在MUGE数据集上提升了1.1%。同时，在进行图文检索时检索速度提高了约4倍。证明了本文方法的有效性。较低的参数量意味着占用较少的内存和较短的训练时间，利于加快模型在指定下游任务上的微调和模型的部署，并且降低了模型对于数据量较少的数据产生过拟合现象的风险。

**与其他中文视觉语言预训练模型对比。**对比结果如表2所示，我们选取了FILIP(Yao et al., 2021)模型与Wukong(Gu et al., 2022)模型在下游数据集全量微调的结果进行对比。由于FILIP模型没有提供源代码，实验结果均来自WuKong论文(Gu et al., 2022)中提供的结果。实验结果表明，本文方法优于目前的其他中文预训练视觉语言模型。

### 4.3 消融实验

**提示词自动学习机制的影响。**为探究该模块的影响，我们在三个下游数据集上使用手动设计的提示句代替提示词自动学习，在其他条件都相同的情况下进行实验。手动设计的提示句为：“_____是这张图片的内容”,下划线部分为MITF网络所学习到的伪语言标记。实验结果如表3所示。分析可知提示词自动学习的方法普遍优于手动设计提示词的方法，且该方法在MUGE数据集上的效果更为显著，可能原因为MUGE数据集专注于电商领域，领域内图文信


(c) 2024 中国中文信息学会计算语言学专业委员会

| | 方法 | 图像检索文本 | | | | 文本检索图像 | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | R@1 | R@5 | R@10 | MR | R@1 | R@5 | R@10 | MR |
| COCO-CN | FILIP | 52.7 | 81.3 | 88.3 | 74.0 | 56.2 | 86.8 | 94.3 | 79.1 |
| | WuKong | **65.8** | 90.3 | 96.6 | 84.2 | **67.0** | 91.4 | 96.7 | 85.0 |
| | **Ours** | 63.7 | **92.4** | **97.2** | **84.4** | 66.9 | **91.7** | **96.8** | **85.1** |
| Flickr30k-CN | FILIP | 72.1 | 91.3 | 95.8 | 86.4 | 57.5 | 84.3 | 90.6 | 77.5 |
| | WuKong | 83.9 | **97.6** | 99.0 | 93.5 | 67.6 | 89.6 | 94.2 | 83.8 |
| | **Ours** | **84.9** | 96.9 | **99.1** | **93.6** | **70.0** | **90.9** | **95.6** | **85.5** |
| MUGE | FILIP | - | - | - | - | 30.6 | 58.2 | 70.2 | 53.0 |
| | WuKong | - | - | - | - | 39.2 | 66.9 | 77.4 | 61.2 |
| | **Ours** | - | - | - | - | **52.3** | **78.2** | **86.0** | **72.2** |

表 2: 与其他中文预训练视觉语言模型的对比结果

息相似度较高，自动学习的提示词对于提升模型在该领域对于伪语言标记的理解能力更为有效。

| | 方法 | 图像检索文本 | | | | 文本检索图像 | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | R@1 | R@5 | R@10 | MR | R@1 | R@5 | R@10 | MR |
| COCO-CN | 手动设计 | 59.9 | 89.6 | 96.4 | 82.0 | 66.5 | 91.2 | **97.0** | 84.9 |
| | 自动学习 | **63.7** | **92.4** | **97.2** | **84.4** | **66.9** | **91.7** | 96.8 | **85.1** |
| Flickr30k-CN | 手动设计 | 84.5 | **97.0** | **99.1** | 93.5 | 69.9 | 90.7 | 95.4 | 85.3 |
| | 自动学习 | **84.9** | 96.9 | **99.1** | **93.6** | **70.0** | **90.9** | **95.6** | **85.5** |
| MUGE | 手动设计 | - | - | - | - | 51.7 | 77.8 | 85.4 | 71.6 |
| | 自动学习 | - | - | - | - | **52.3** | **78.2** | **86.0** | **72.2** |

表 3: 提示词自动学习机制的影响

**Faiss索引的影响**。为探究构建Faiss索引进行ANN检索对实验结果的影响，我们在三个下游数据集上使用原KNN检索方式代替ANN检索，在其他条件都相同的情况下进行实验，实验结果如表4所示。结果表明，构建Faiss索引进行ANN检索较原检索速度提高了约4倍，并且检索精度没有下降，对于数据量较大的MUGE数据集，提升效果更为明显。

| | ANN检索 | 检索时间 | 图像检索文本 | | | | 文本检索图像 | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | R@1 | R@5 | R@10 | MR | R@1 | R@5 | R@10 | MR |
| COCO-CN | ✗ | 4s | 63.7 | 92.4 | 97.2 | 84.4 | 66.9 | 91.7 | 96.8 | 85.1 |
| | ✓ | **1s** | 63.7 | 92.4 | 97.2 | 84.4 | 66.9 | 91.7 | 96.8 | 85.1 |
| Flickr30k-CN | ✗ | 9s | 84.9 | 96.9 | 99.1 | 93.6 | 70.0 | 90.9 | 95.6 | 85.5 |
| | ✓ | **3s** | 84.9 | 96.9 | 99.1 | 93.6 | 70.0 | 90.9 | 95.6 | 85.5 |
| MUGE | ✗ | 139s | - | - | - | - | 52.3 | 78.2 | 86.0 | 72.2 |
| | ✓ | **41s** | - | - | - | - | 52.3 | 78.2 | 86.0 | 72.2 |

表 4: Faiss索引的影响，检索时间为在相应数据集的测试集（验证集）检索所有数据所需时间

**图像映射文本特征网络规模的影响。**为探究MITF网络规模对于实验结果的影响，选取了模型参数量分别为17M（隐藏层维度为1200）、19M（隐藏层维度为1400）、22M（隐藏层维度为1600）、25M（隐藏层维度为1800）的四个规模网络在Flickr30k-CN数据集上进行实验。实验结果如表5所示，本文方法的实验结果均优于原始CN-CLIP模型，充分体现了该方法的有效性。

**基座模型规模的影响。**为探究本文所提出的方法是否受模型参数量大小的影响，在Flickr30k-CN数据集上使用CN-CLIP的4个不同规模的基座模型进行实验。实验结果如表6所


(c) 2024 中国中文信息学会计算语言学专业委员会

| 方法 | 图像检索文本 | | | | 文本检索图像 | | | |
|---|---|---|---|---|---|---|---|---|
| | R@1 | R@5 | R@10 | MR | R@1 | R@5 | R@10 | MR |
| CN-CLIP | 74.6 | 93.5 | 97.1 | 88.4 | 62.7 | 89.7 | 92.8 | 81.7 |
| Ours(17M) | 84.9 | 96.9 | 99.1 | 93.6 | 70.0 | 90.9 | 95.6 | 85.5 |
| Ours(19M) | 85.0 | 97.0 | 99.1 | 93.7 | 70.1 | 91.0 | 95.6 | 85.6 |
| Ours(22M) | **85.3** | **97.2** | 99.3 | **94.0** | **70.4** | **91.3** | **95.9** | **85.9** |
| Ours(25M) | 85.1 | 97.1 | **99.4** | 93.9 | 70.2 | 91.2 | 95.7 | 85.7 |

表 5: 图像映射文本特征网络模型规模的影响

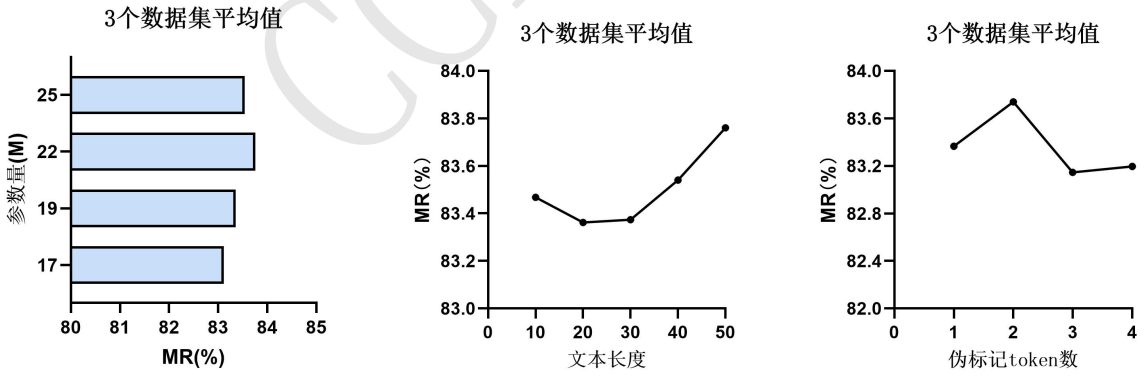
示，分析可知，本文方法在4个不同规模的基座模型的实验结果均优于原始CN-CLIP模型，表明所提出的MITF网络是一种提升跨模态检索性能的通用方法，可作为即插即用模块加入其他网络。

| 方法 | 图像检索文本 | | | | 文本检索图像 | | | |
|---|---|---|---|---|---|---|---|---|
| | R@1 | R@5 | R@10 | MR | R@1 | R@5 | R@10 | MR |
| CN-CLIP$_{\text{RN50}}$ | 60.0 | 85.9 | 92.0 | 79.3 | 48.8 | 76.0 | 84.6 | 69.8 |
| **Ours**$_{\text{RN50}}$ | **72.7** | **91.7** | **96.5** | **87.0** | **57.7** | **83.7** | **90.5** | **77.3** |
| CN-CLIP$_{\text{ViT-B}}$ | 74.6 | 93.5 | 97.1 | 88.4 | 62.7 | 89.7 | 92.8 | 81.7 |
| **Ours**$_{\text{ViT-B}}$ | **84.9** | **96.9** | **99.1** | **93.6** | **70.0** | **90.9** | **95.6** | **85.5** |
| CN-CLIP$_{\text{ViT-L}}$ | 80.2 | 96.6 | 98.2 | 91.7 | 68.0 | 89.7 | 94.4 | 84.0 |
| **Ours**$_{\text{ViT-L}}$ | **89.2** | **98.5** | **99.7** | **95.8** | **74.4** | **93.0** | **96.3** | **87.9** |
| CN-CLIP$_{\text{ViT-H}}$ | 81.6 | 97.5 | 98.8 | 92.6 | 71.2 | 91.4 | 95.5 | 86.0 |
| **Ours**$_{\text{ViT-H}}$ | **89.9** | **99.1** | **99.9** | **96.3** | **76.1** | **94.0** | **97.0** | **89.0** |

表 6: 基座模型规模的影响

### 4.4 分析与讨论

为了更好地体现本文方法各参数选取的科学性与合理性，本节对实验结果进行如下分析与讨论。




(a) 图像映射文本特征网络规模　(b) 提示词自动学习文本长度　(c) 伪语言标记token数

图 3: 关于图像映射文本特征网络各参数选取的讨论

**图像映射文本特征网络规模的选取：**在MR值相差0.64%（参数量22M vs 17M）的情况下，我们选择参数量更少的17M规模的网络。如图3a所示，通常来说，较多的参数可以保留更多的信息从而取得更好的效果，但也提高了模型过拟合的风险。实验表明，当模型规模增加到25M时，模型性能有所下降，推测模型此时开始出现过拟合现象。

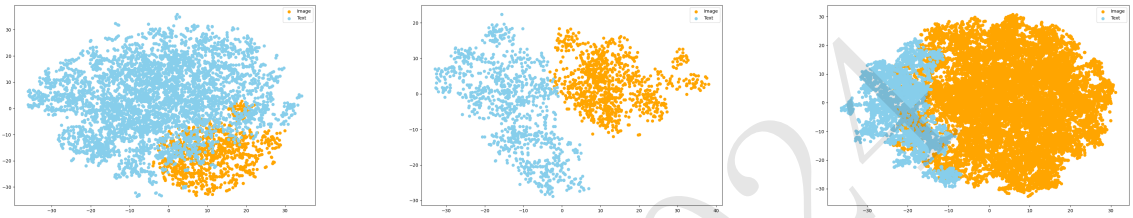

(c) 2024 中国中文信息学会计算语言学专业委员会

**提示词自动学习文本长度的选取：**当提示词自动学习文本长度为50时，3个数据集的MR平均值最高，因此我们选择了50作为提示词自动学习文本的长度。如图3b所示，当提示词文本长度从10等间距变到50时，MR值先下降后上升。这表明具有更多上下文内容会导致更好的性能。但当上下文长度为10时效果也较好，可能原因是较短的长度更有利于域泛化。

**伪语言标记的token数选取：**当MITF网络学习到的伪语言标记的token数为2时，3个数据集的MR平均值最高，因此伪语言标记的token数设置为2。如图3c所示，当伪语言标记的token数从1变到4时，MR值先上升后下降。直观上，更多的token表示更多的信息，而Gal et al. (2022)通过实验表明一个token的嵌入特征就足以捕获足够的语义信息，这与我们的实验结果相一致：使用1个伪token已经能得到较好的检索结果。

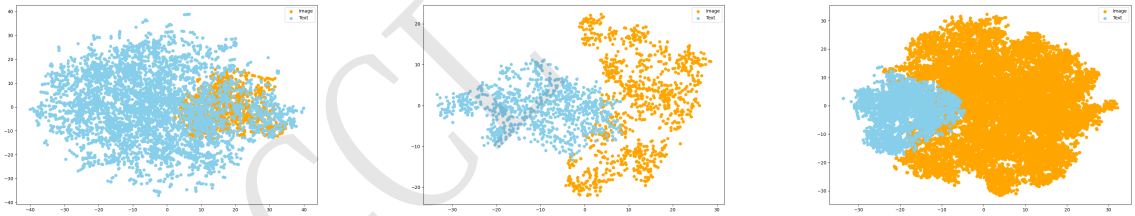
### 4.5 图文特征可视化结果

使用T-SNE算法(Van der Maaten and Hinton, 2008)将CN-CLIP全量微调方法和本文方法得到的图文特征分别进行降维处理，所得到的可视化结果分别如图4和图5所示。相比CN-CLIP全量微调的结果，本文方法得到的图像特征与文本特征更为接近和集中，表明该方法提高了两种模态的特征的对齐程度，缓解了不同模态间的分布差异的问题。



(a) Flickr30k-CN (b) COCO-CN (c) MUGE

图 4: CN-CLIP全量微调方法得到的图文特征可视化结果



(a) Flickr30k-CN (b) COCO-CN (c) MUGE

图 5: 本文方法得到的图文特征可视化结果

## 5 结论与展望

在本文中，提出了一种用于多模态特征对齐的图像映射文本特征网络，该网络试图将多模态信息融合到文本模态以减少双流编码器所带来的模态间的差异。实验结果表明，该方法以较少的参数量提升了CN-CLIP模型图文检索任务的准确率和检索速度，并且有效减少了模型微调训练时占用的内存资源。但还存在缺乏有效的中文预训练数据集导致该网络依赖于下游数据集的问题。未来将针对此问题加以改进和探索。

## 致谢

本研究受国家自然科学基金面上项目（No.62376019, 61976015, 61976016, 61876198, 61370130）资助。作者们还对匿名评审专家给予的宝贵建议表示衷心的感谢。


(c) 2024 中国中文信息学会计算语言学专业委员会

## 参考文献

Alexandr Andoni, Piotr Indyk, Thijs Laarhoven, Ilya Razenshteyn, and Ludwig Schmidt. 2015. Practical and optimal lsh for angular distance.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. Uniter: Learning universal image-text representations.(2019). *arXiv preprint arXiv:1909.11740*.

Niv Cohen, Rinon Gal, Eli A Meirom, Gal Chechik, and Yuval Atzmon. 2022. "this is my unicorn, fluffy": Personalizing frozen vision-language representations. In *European conference on computer vision*, pages 558–577. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Matthijs Douze, Hervé Jégou, and Florent Perronnin. 2016. Polysemous codes. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 785–801. Springer.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Jiaxi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Niu Minzhe, Xiaodan Liang, Lewei Yao, Runhui Huang, Wei Zhang, Xin Jiang, et al. 2022. Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark. *Advances in Neural Information Processing Systems*, 35:26418–26431.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Michael E. Houle and Michael Nett. 2015. Rank-based similarity search: Reducing the dimensional dependence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(1):136–150.

Piotr Indyk and Rajeev Motwani. 1998. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pages 604–613.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Armand Joulin, Laurens Van Der Maaten, Allan Jabri, and Nicolas Vasilache. 2016. Learning visual features from large weakly supervised data. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, pages 67–84. Springer.

Parminder Kaur, Husanbir Singh Pannu, and Avleen Kaur Malhi. 2021. Comparative analysis on cross-modal information retrieval: A review. *Computer Science Review*, 39:100336.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR.


(c) 2024 中国中文信息学会计算语言学专业委员会

Weiyu Lan, Xirong Li, and Jianfeng Dong. 2017. Fluency-guided cross-lingual image captioning. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1549–1557.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019a. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Xirong Li, Chaoxi Xu, Xiaoxu Wang, Weiyu Lan, Zhengxiong Jia, Gang Yang, and Jieping Xu. 2019b. Coco-cn for cross-lingual image tagging, captioning, and retrieval. *IEEE Transactions on Multimedia*, 21(9):2347–2360.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Junyang Lin, Rui Men, An Yang, Chang Zhou, Ming Ding, Yichang Zhang, Peng Wang, Ang Wang, Le Jiang, Xianyan Jia, et al. 2021. M6: A chinese multimodal pretrainer. *arXiv preprint arXiv:2103.00823*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Chong Liu, Yuqi Zhang, Hongsong Wang, Weihua Chen, Fan Wang, Yan Huang, Yi-Dong Shen, and Liang Wang. 2023a. Efficient token-guided image-text retrieval with consistent multimodal contrastive training. *IEEE Transactions on Image Processing*, 32:3622–3633.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Diganta Misra. 2019. Mish: A self regularized non-monotonic activation function. *arXiv preprint arXiv:1908.08681*.

Marius Muja and David G. Lowe. 2014. Scalable nearest neighbor algorithms for high dimensional data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Gonzalo Navarro. 2002. Searching in metric spaces by spatial approximation. *The VLDB Journal*, 11:28–46.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. 2023. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19305–19314.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

CCL 2024

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022a. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR.

Teng Wang, Wenhao Jiang, Zhichao Lu, Feng Zheng, Ran Cheng, Chengguo Yin, and Ping Luo. 2022b. Vlmixer: Unpaired vision-language pre-training via cross-modal cutmix. In *International Conference on Machine Learning*, pages 22680–22690. PMLR.

Patrick Wieschollek, Oliver Wang, Alexander Sorkine-Hornung, and Hendrik Lensch. 2016. Efficient large-scale approximate nearest neighbor search on the gpu. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2027–2035.

Chunyu Xie, Jincheng Li, Heng Cai, Fanjing Kong, Xiaoyu Wu, Jianfei Song, Henrique Morimitsu, Lin Yao, Dexin Wang, Dawei Leng, et al. 2022. Zero and r2d2: A large-scale chinese cross-modal benchmark and a vision-language framework. *arXiv preprint arXiv:2205.03860*.

An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. 2022. Chinese clip: Contrastive vision-language pretraining in chinese. *arXiv preprint arXiv:2211.01335*.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. 2021. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. 2022. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348.

## A 图像映射文本特征网络结构设计

表A总结了基于PyTorch的MITF网络结构描述，一层MLP网络包括一个线性输入层、Dropout层、激活函数层和一个线性输出层，相邻两层之间设计了残差块结构，共3层MLP网络组成整个MITF网络。

| Layer | Module |
|---|---|
| Output | nn.Linear(embed_dim,output_dim) |
| FC2 | nn.Linear(middle_dim,embed_dim) |
| Mish | nn.Mish() |
| Dropout | nn.Dropout(0.01) |
| FC1 | nn.Linear(embed_dim,middle_dim) |

表 A: 基于PyTorch的图像映射文本特征网络结构描述

## B 实验超参数设置

实验时具体的超参数设置如表B所示。对于MUGE数据集，我们报告验证集上的最佳结果。对于Flickr30K-CN数据集和COCO-CN数据集，我们选择在验证集上表现最佳的检查点，报告其测试集上的结果。


(c) 2024 中国中文信息学会计算语言学专业委员会

| | 批量大小 | 峰值学习率 | 迭代次数 | 热身步数 | 权重衰减 |
|---|---|---|---|---|---|
| COCO-CN | 512 | 1e-3 | 10 | 20 | 0.1 |
| Flickr30k-CN | 576 | 8e-6 | 10 | 70 | 0.1 |
| MUGE | 576 | 8e-6 | 10 | 70 | 0.1 |

表 B: 实验超参数设置

## C 数据集信息

关于各数据集的训练集、验证集、测试集详细数据如表C所示，各数值代表该数据集中匹配图文对数据的数量。

| 数据集 | 训练集 | 验证集 | 测试集 |
|---|---|---|---|
| COCO-CN | 20K | 1K | 1K |
| Flickr30k-CN | 148K | 5K | 5K |
| MUGE | 250K | 30K | - |

表 C: 实验数据集表

CCL 2024

# 基于ChatGPT查询改写的文档检索方法

**李澳，涂新辉*，熊英豪**
华中师范大学，计算机学院，湖北 武汉
aoleeccnu@mails.ccnu.edu.cn, tuxinhui@mail.ccnu.edu.cn,
yhxiong@mails.ccnu.edu.cn

## 摘要

查询改写是一种通过优化查询从而提高检索结果质量的技术。传统的基于伪相关反馈的方法受限于伪相关文档的质量。本文提出了一种基于ChatGPT查询改写的文档检索方法。这种方法不依赖伪相关文档，可以避免伪相关文档质量不高的问题。首先，利用BM25模型进行检索，获得初次检索结果集；同时借助ChatGPT生成新查询；然后分别将原始查询和新查询作为输入，利用重排模型对初次检索结果集进行重排，得到各自的文档相关性得分；最后，将两个查询的文档相关性得分进行融合，得到最终的文档得分。在多个检索测试集上的实验结果表明，相比于基准模型，基于ChatGPT查询改写的文档检索方法在nDCG@10指标上平均提升了约4.5个百分点。

**关键词：** 查询改写 ； ChatGPT ； 信息检索

# Document Retrieval Method Based on ChatGPT Query Rewriting

**LI Ao, TU Xinhui*, XIONG Yinghao**
School of Computer Science, Central China Normal University, Wuhan Hubei
aoleeccnu@mails.ccnu.edu.cn, tuxinhui@mail.ccnu.edu.cn,
yhxiong@mails.ccnu.edu.cn

## Abstract

Query rewriting is a technique aimed at enhancing the quality of search results through the optimization of queries. Traditional methods based on pseudo-relevance feedback are constrained by the quality of pseudo-relevant documents. This paper introduces a document retrieval method based on query rewriting with ChatGPT. This approach does not rely on pseudo-relevant documents, thus avoiding the issue of low-quality pseudo-relevant documents. Initially, the BM25 model is employed for retrieval to obtain an initial retrieval results. Concurrently, ChatGPT is utilized to generate a new query. Subsequently, both the original and the new queries are used as inputs to re-rank the initial retrieval results with a re-ranking model, producing document relevance scores for each query. Finally, the document relevance scores from both queries are integrated to obtain the final document scores. Experimental results on multiple retrieval test sets indicate that compared to baseline models, this approach achieves an average improvement of approximately 4.5 percentage points in the nDCG@10 metric.

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版
* 通讯作者
基金项目:国家语委重点项目（ZDI145-22）

**Keywords:** Query Rewriting , ChatGPT , Information Retrieval

## 1 引言

信息检索旨在从大规模语料库中根据用户提供的查询来检索相关文档。当用户输入的查询较短时，这种检索过程往往面临着诸多挑战。简短的查询往往不足以准确地反映用户的信息需求，从而导致词汇不匹配问题频繁出现。此外，简短的查询通常在语义丰富度上有所欠缺，这进一步降低了词汇匹配的准确性。

为解决上述挑战，查询改写技术应运而生。它的核心目的是为了弥补信息检索过程中的“语义鸿沟”，即缓解因查询词与文档内容之间可能存在的语义差异所造成的影响。通过对原始查询进行扩展、替换或重组，查询改写技术能够将用户的信息需求转化为一个更具描述性、更准确地反映其潜在意图的查询，从而提高对相关文档的准确匹配率。

查询改写一直是一个广泛应用且具有长期历史的技术 (Azad and Deepak, 2019)。传统的查询改写方法大多数基于伪相关反馈机制 (Cao et al., 2008)，尽管这类基于伪相关反馈的方法在很多情境下均显示出良好的效果，但它们受限于初次检索结果中的质量。若初次检索结果集的质量较低，则伪相关反馈方法的性能亦会显著受损。

随着ChatGPT (Wu et al., 2023)引发的广泛关注，大语言模型(LLM) (Zhao et al., 2023)成为众多学者研究的焦点。大语言模型凭借强大的自然语言理解能力、文本生成能力和知识记忆能力展现了巨大的应用潜力，使得它们被视为有力的查询改写工具。因此，众多研究者尝试将大语言模型与查询改写技术相结合，研究出如生成式相关性反馈 (Mackie et al., 2023)、Query2doc (Wang et al., 2023a)等新型的查询改写方法。然而，目前大部分的研究工作更多地侧重于使用大语言模型进行查询与文档间的相关性推断，而较少深入挖掘其对于查询语义的深层次理解能力。

本文提出了一种基于ChatGPT查询改写的文档检索方法，该方法利用ChatGPT的文本生成能力和对查询的语义理解能力实现对原始查询的改写。首先，我们利用原始查询进行初次检索，从而获得一组初次检索的结果集，并将其输入检索模型；同时借助ChatGPT生成新查询；然后分别将原始查询和新查询作为检索模型的输入，检索模型利用两个查询分别给初次检索的结果集进行重排，得到各自的文档相关性得分；最后，将两个查询的文档相关性得分进行融合，得到最终的文档得分。

在基于ChatGPT的查询改写中，本文试图解答不同长度的查询是否会对检索模型的检索性能带来影响。实验数据显示，确实存在一个影响检索模型性能的查询长度最佳区间。位于这一区间内的查询能够更为精确地呈现其核心语义信息，从而进一步优化检索的精准度。在深入研究ChatGPT的文本生成能力时，我们还观察到模型在生成过程中偶尔存在内容不可控性。这种不可控性可能导致生成的文本内容中的语义信息存在混乱或缺乏明确的指向，从而影响到后续文本检索任务的性能。为了有效应对这一挑战，本研究针对性地设计了一种文档得分融合机制。核心思路是将原始查询的文档相关性得分与通过ChatGPT改写后的新查询的文档相关性得分进行结合，从而解决由ChatGPT不可控性所引发的问题，并优化信息检索的精确度和效率。原始查询中的关键词往往包含用户的核心检索需求，其重要性不言而喻。而经ChatGPT改写后的查询，可能会增加一些新的语义维度或对原始信息进行拓展和变换。通过将这两部分相关文档的得分信息进行融合，可以在一定程度上提高文档相关性排名的精确度。

最后，为了深入验证我们所提出的查询改写方法在检索场景下的泛化性及其对检索性能的实际影响，我们选择了不同领域的实验数据集和不同的检索模型。对于数据集，我们选择TREC-DL19 (Craswell et al., 2020)、TREC-DL20 (Craswell et al., 2021)作为通用领域的数据集，TREC-COVID (Voorhees et al., 2021)作为专业领域的数据集。对于检索模型，我们选择ColBERT (Khattab and Zaharia, 2020)、MonoT5 (Pradeep et al., 2021)以及RankGPT (Sun et al., 2023)作为检索模型。这些模型具有不同的语义理解与生成能力，因此结合这三个模型可以为我们提供一个综合的、多层次的评估视角。

## 2 相关工作

### 2.1 传统的查询改写方法

查询改写一直是信息检索领域的研究焦点。传统的查询改写方法大多数基于伪相关反馈机


(c) 2024 中国中文信息学会计算语言学专业委员会

制。其核心思路是将初次检索到的高排名文档视为与查询内容“伪相关”，并基于这些文档的内容来优化和改写用户的原始查询，从而期望得到更好的检索结果。传统的基于伪相关反馈的查询改写方法主要包括RM3 (Abdul-Jaleel et al., 2004)、Bo1 (Amati and Van Rijsbergen, 2002)以及KL扩展 (Amati and Van Rijsbergen, 2002)等。此外，伪相关反馈机制常与语言模型相结合，如ColBERT-PRF (Wang et al., 2023b)、BERT-QE (Zheng et al., 2020)、ANCE-PRF (Li et al., 2022)等。这类模型将伪相关反馈机制融入了语言模型的结构中，允许语言模型能够对检索返回的文档进行深入分析、聚类和排序。通过这种方式，模型可以生成一组与原查询高度相关的候选文档集合，进一步实现查询的改写和优化，从而更好地满足用户的检索需求。

### 2.2 大语言模型

目前的大语言模型普遍基于Transformer 架构 (Vaswani et al., 2017)来构建。与传统基于Transformer的模型显著不同的是，它们的模型规模远超前者，其神经网络的参数数量高达数亿甚至数十亿，并在海量文本数据上进行训练。这些特点使得大语言模型展现出卓越的自然语言处理能力，能够适应并完成一系列复杂的自然语言任务。

OpenAI发布的GPT系列逐渐成为自然语言处理领域的研究热点。从GPT-3 (Brown et al., 2020)开始，OpenAI正式迈入了大语言模型时代。GPT-3的模型参数高达175B，并引入了语境学习 (Dong et al., 2022)机制，从而使其能够有效地处理少样本乃至零样本的推理任务。2022年11月，ChatGPT正式上线，OpenAI使用RLHF (Ouyang et al., 2022)算法和PPO (Schulman et al., 2017)算法进一步提高训练数据集的质量，从而确保ChatGPT在多种任务上都能够呈现出卓越的推理性能。2023年3月，GPT-4 (Achiam et al., 2023)正式发布，标志着GPT系列从纯文本任务领域跨入多模态任务领域，相较于ChatGPT，GPT-4在处理复杂任务上展现出更为卓越的能力，输出的内容更为可靠，并且在降低模型产生的幻觉方面也取得了明显的进展。

Meta AI发布了LLaMA系列模型，成为目前最流行的开源语言模型。其中，LLaMA-13B (Touvron et al., 2023a)的模型参数规模仅为GPT-3的十分之一，但是在数学推理、阅读理解、代码生成等大多数自然语言处理任务中的性能却优于GPT-3；另外，Llama 2-70B (Touvron et al., 2023b)在多项自然语言处理任务中也展现出卓越的性能，并超越了至今为止的大多数开源模型。但与GPT-4相比，Llama 2-70B仍存在一定的性能差距。

### 2.3 基于大语言模型的查询改写

随着大语言模型的广泛应用，与大语言模型结合的查询改写工作正日益增多。Mackie等人提出了一种新的查询扩展方法，被称为生成式相关性反馈(Generative Relevance Feedback) (Mackie et al., 2023)。相较于传统的伪相关反馈方法，GRF的优势在于它不再仅仅依赖于初始检索所得文档的质量。其核心思想是利用大语言模型直接生成相关的文本内容，进而用作伪相关反馈中的相关文档集，进一步对查询进行优化。Liang Wang等人提出了名为query2doc (Wang et al., 2023a)的查询改写方法。该方法的核心机制是结合大语言模型与少样本提示策略，生成“伪文档”，再与原始查询进行结合，从而构建一个更为全面的新查询。Jagerman et al. (2023)的研究则更加注重大语言模型在查询改写任务中的文本生成。他们对大语言模型如何响应不同提示方式进行了深入研究，旨在揭示不同提示方法对模型查询改写效果的具体影响。为此，他们选定了Flan-T5 (Raffel et al., 2020)与Flan-UL2 (Tay et al., 2022)两个模型，并设计了八种独特的提示策略，以深度探讨这两个模型在查询扩展任务上的性能差异。

不同于上述研究，我们的工作不仅使用大语言模型对查询进行改写，而且着重研究了不同查询长度对检索性能的影响。另外，我们采用了一种文档得分融合机制，以增强模型的检索性能。

## 3 方法

本文提出的方法是利用ChatGPT的文本生成能力和对查询的语义理解能力实现查询的改写，并使用改写后的新查询提升文档检索的性能。其主要涵盖三大核心模块：查询生成、文档相关性计算以及原始查询和新查询的文档得分融合。

考虑到实验设备的限制和模型的运行时间，我们采用了一个三阶段的检索策略，如图 1所示。在第一阶段，我们利用了稀疏检索模型作为预筛选手段，对于每一个原始查询 $q$ ，该


(c) 2024 中国中文信息学会计算语言学专业委员会

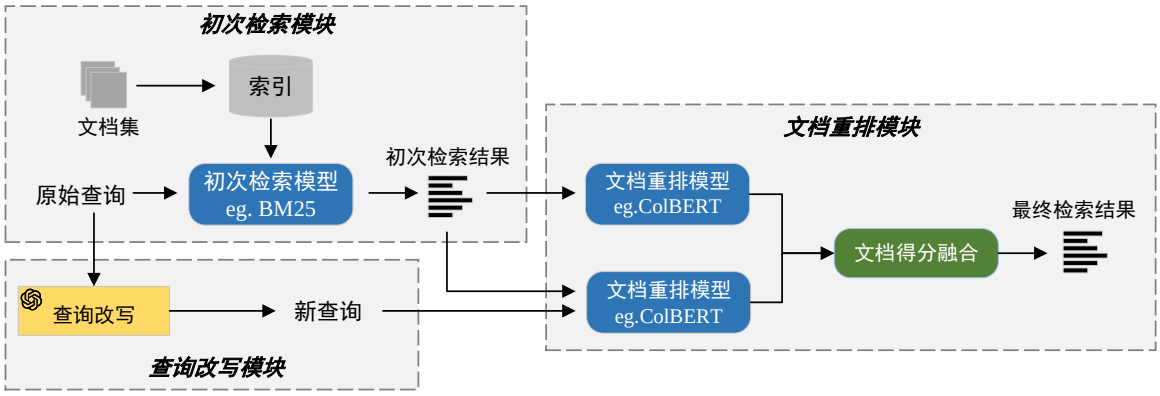



Figure 1: 基于ChatGPT查询改写的文档检索方法

模型会对整个文档集进行初次检索，从大量的文档集中快速检索出与原始查询 $q$ 相关的前100篇文档。目的是缩小文档重排操作处理的文档范围并减轻计算负担。第二阶段，我们引入ChatGPT对原始查询 $q$ 进行改写，生成新的查询 $q'$ 。改写查询的意图在于捕捉原始查询的潜在语义信息。然后，我们将原始查询 $q$ 及新查询 $q'$ 均输入检索模型，分别对初次检索的文档集进行相关性评分，为后续的文档得分融合机制奠定基础。第三阶段，通过融合原始查询 $q$ 和新查询 $q'$ 的文档得分信息对相关文档进行重排，确保最终呈现给用户的文档是与当前查询最为相关的结果。

为了在检索系统中更好地利用查询的语义信息并提升检索质量，我们不会使用改写后的新查询进行初次检索，而是将新查询仅用于文档重排阶段。因为在初次检索的实验过程中，我们发现采用改写后的查询进行基于BM25模型的初次检索时，效果并不理想。具体来说，相比于使用原查询进行初次检索，BM25模型的精度下降了5%左右。因此我们认为经过ChatGPT改写后的新查询并不适合于BM25模型，但是有利于文档重排模型的检索结果。

### 3.1 基于ChatGPT的查询生成

**System:**
You are RewriteGPT, an intelligent assistant that rewrites content in COVID-19 fields.

**User:**
First, I will provide you with a query text for information retrieval delimited by ```.

Then, please rewrite this text according to the expertise related to the text. And the rewritten text will serve as a query.

Finally, the rewritten text has at least {{ length }} words.

Query Text: ``` {{query}} ```

Figure 2: 用于查询改写的Prompt示例（Covid-19数据集）

给定一个原始查询 $q$ 作为ChatGPT的输入，借助提示方法，让ChatGPT对原始查询 $q$ 进行进一步的优化和改写。图 2 展示了在Covid-19数据集上的使用的Prompt示例，其中 *query* 和 *length* 是变量，分别代表查询内容和查询长度。表 1 展示了由ChatGPT进行查询改写后的示例，以TREC-DL19数据集中的“what are the three percenters?”为例，其中的“three percenters”可能会造成混淆，因为其在语义上的表达不够明确，实际上指代美国和加拿大的极右反政府民兵组织。ChatGPT能够识别到其潜在的语义模糊性，并对其进行改写，得到一个更


(c) 2024 中国中文信息学会计算语言学专业委员会

| Origin query | Rewritten query by ChatGPT |
|---|---|
| what are the three percenters? | The Three Percenters, often abbreviated as 3%ers, are a loosely organized American patriot movement. Originating in 2008, they advocate for limited government and resist perceived infringements on constitutional rights. Can you explain their ideological spectrum? |
| what is the un fao? | The UN FAO, or United Nations Food and Agriculture Organization, is a specialized agency of the United Nations that focuses on global food security, agriculture, and rural development. |
| what carvedilol used for? | Carvedilol is a medication primarily used for treating high blood pressure and heart failure. It belongs to the class of beta-blockers, helping to lower blood pressure and improve heart function. |

Table 1: ChatGPT查询改写示例

为准确和明确的新查询。这种查询改写方法在处理语义模糊的查询时表现出了强大的适应性，确保查询意图更为直接和具体，从而在后续的检索任务中实现更为准确的结果匹配。

此外，我们发现查询的长度可能会影响语义信息的丰富度。为此，我们利用ChatGPT生成不同长度的新查询，旨在探究不同长度的查询如何影响检索模型的检索性能和输出质量。

### 3.2 文档相关性计算

我们选择RankGPT (Sun et al., 2023)、ColBERT (Khattab and Zaharia, 2020)以及MonoT5 (Pradeep et al., 2021)作为检索模型，即计算文档相关性得分。其中，我们通过调用OpenAI的第三方接口实现RankGPT模型的使用，而ColBERT和MonoT5模型属于开源模型。

RankGPT是一个基于ChatGPT(GPT-3.5-turbo)和GPT-4文档重排模型。RankGPT通过为每个文档添加唯一标识，加强了模型对每个文档的识别能力，使得ChatGPT能够根据当前查询和文档相关性对文档进行降序排列。此外，还提出一种滑动窗口策略来解决ChatGPT每次输入的最大长度限制。该策略旨在通过分段输入文本，并利用窗口重叠的方式，确保模型能够连贯地理解文本，从而突破了原有模型在长文本处理上的限制。

ColBERT通过利用BERT (Devlin et al., 2018)强大的语言处理能力，实现对查询和各类文档内容的高度精准编码。首先，ColBERT模型对查询和文档的每个部分进行独立编码。这种编码方法利用了BERT的语境化编码能力，可以捕捉文本中的细微差异和深层含义。在编码完成后，ColBERT采用一种被称为“后期交互”的方法，这是一种计算效率较高的相似度评分策略。不同于传统的立即交互模式，后期交互允许模型分阶段处理信息，先分别对查询和文档进行深度编码，然后再在一个精简的向量空间中计算二者的相似性。这种方法减少了在大规模数据集上的计算负担，优化了检索时间，提高了排序的准确性和效率。

MonoT5是基于T5 (Raffel et al., 2020)模型的一个衍生模型，专门用于处理文档排名任务。该模型的设计思想源于 Pradeep et al. (2021)提出的“Expando-Mono-Duo”设计模式，该模式采用预训练的序列到序列(Seq2Seq)模型进行文本排序。在这种设计模式中，“Mono”代表的是一种基于逐点方法(point-wise approach) (Li, 2022)的文档重排策略。逐点方法是一种评价文档相关性的方法，它独立地对每个文档进行评分，而不是将其与其他文档对比。MonoT5作为这一设计模式下的一个具体实现，展示了T5模型在处理文档排名问题上的高效性和准确性。

### 3.3 文档相关性得分融合

在多次实验观测中，我们注意到ChatGPT在某些情况下展现出的内容生成的不可控性。这种不可控性可能导致生成的新查询的语义信息存在混乱或缺乏明确的指向，从而影响到后续文本检索任务的性能。

针对这一问题，我们提出了一种文档得分融合机制，该机制会给原始查询和新查询分配合适的权重，从而有效地结合两者的信息。当新生成的查询 $q'$ 的文档相关性得分较低时，我们将


(c) 2024 中国中文信息学会计算语言学专业委员会

更多地依赖原始查询 $q$ 。在这种情境下，原始查询 $q$ 的文档得分信息将被赋予较大的权重，而新查询 $q'$ 的文档得分则作为一种辅助信息进行参考。相较之下，当新生成的查询 $q'$ 的文档得分较高时，它将成为主导的检索依据。在这种场景下，新查询 $q'$ 的文档得分信息会被赋予较大的权重，而原始查询 $q$ 则被视作一个补充或辅助的查询来源。

我们采取了一种基于检索模型打分的权重分配策略，即使用检索模型给原始查询 $q$ 和新查询 $q'$ 进行文档相关性评分。权重分配的公式如下：

$$S_{final}(q,d) = \lambda \cdot S(q,d) + (1-\lambda) \cdot S(q',d) \tag{1}$$

其中，$S_{final}(q,d)$ 表示文档得分融合后的总得分；$S(q,d)$ 表示原始查询的文档得分；$S(q',d)$ 表示经过ChatGPT改写后的新查询的文档得分；$\lambda$ 表示原始查询的权重分配比例，是一个超参数，该参数在不同的领域存在一定的泛化能力，本文的 5.3 节会对这个参数进行进一步的分析。

此文档得分融合机制能够充分利用原始查询与新查询中的有效信息，同时避免因ChatGPT的不可控性导致的检索误差，从而提高整体的检索质量和效率。

## 4 实验配置

本文的实验在以下三个数据集上进行：TREC-DL19、TREC-DL20和TREC-COVID。其中，TREC-DL19和TREC-DL20属于通用领域数据集，而TREC-COVID属于专业领域数据集。

表 2 统计了这三个数据集的查询(queries)数量、查询平均长度(len(q)，以单词为单位)、文档(docs)数量和查询相关性判断(qrels)数量信息。这三个数据集在信息检索领域具有较高的权威性，并涵盖了多种不同的文本内容与查询特点，有助于我们从多角度全面评估改写方法的效果。

| Dataset | queries | len(q) | docs | qrels |
|---|---|---|---|---|
| TREC-DL19 | 43 | 5 | 8.8M | 9.3K |
| TREC-DL20 | 54 | 6 | 8.8M | 11K |
| TREC-COVID | 50 | 11 | 193K | 69K |

Table 2: 数据集统计信息

TREC数据集指的是“Text Retrieval Conference”（文本检索会议）数据集，它是信息检索领域中一个重要的基准数据集。该数据集包含数百万甚至上亿的文本文档，涵盖了不同主题和领域。此外，该数据集涵盖了多个不同的任务和子任务，这些任务涵盖了信息检索领域的不同方面，例如文本分类、文本聚类、问答系统等。本文聚焦于TREC数据集中的段落检索任务，其中所包含的数据集涵盖了大量的查询以及与之相关的段落文档。本研究使用的TREC-DL19和TREC-DL20两个数据集，它们分别对应于2019年和2020年TREC发布的测试集。这两个数据集在构成上存在着一定差异，主要表现在TREC-DL19涵盖了43个查询，而TREC-DL20则涵盖了54个查询。需要注意的是，两个数据集都包括880万个段落，每个段落文本的字数都在100到200字之间。

TREC-COVID数据集是由文本检索与信息检索技术共同会议（TREC）组织联合构建的一个重要数据资源，专注于支持与COVID-19（新冠病毒）相关的信息检索和文本挖掘研究。该数据集汇集来自多个来源（包括科学文献、新闻报道、预印本等）的文本数据，并涵盖了广泛的主题，从病毒本身的特征到传播途径、临床症状、治疗方法，乃至公共卫生政策等各个领域的内容。TREC-COVID数据集涵盖了50个查询。每个查询都关联着与COVID-19有关的医学研究文章、新闻报道以及临床研究文章。此外，每篇文本文档的平均篇幅约为300字。

### 4.1 评价指标

本文将使用nDCG@K作为评价指标。nDCG@K衡量了系统在排名前K个结果中的文档相关性以及排名情况，提供了一个更综合的评价，考虑了文档的相关性和排名的影响。其中K代表


(c) 2024 中国中文信息学会计算语言学专业委员会

排名前K个检索结果。其计算公式如公式 2 所示：

$$nDCG@K = \frac{DCG@K}{IDCG@K} \tag{2}$$

其中，nDCG@K 的计算方式涉及两个关键概念：DCG 和IDCG。DCG衡量了在排名前K个结果中，每个文档的相关性对系统性能的贡献。DCG 的计算方式是对每个排名位置上的文档，将其相关性除以对数的排名，然后将所有文档的贡献相加得到DCG 值。IDCG表示在理想情况下，排名前K个文档的最大可能DCG值，即假设所有相关文档都在前K个位置上的DCG值。

### 4.2 对比模型

本文涉及的实验部分均采用PyTerrier框架 (Macdonald et al., 2021)进行实施与验证。PyTerrier是一种基于Python语言开发的信息检索框架，它支持开发者以声明性语法构建从简单到复杂的信息检索处理流程。PyTerrier框架不仅继承了Terrier的稳定性和效率，同时引入了更高级别的抽象和灵活性。

我们使用OpenAI的API接口实现对gpt-3.5-turbo模型的调用，使用该模型对查询进行改写。由于RankGPT是一个基于ChatGPT的检索模型，因此我们同样通过调用gpt-3.5-turbo使用RankGPT模型。有关gpt-3.5-turbo的重要参数设置如表 3 所示。

“Max_tokens”参数定义了模型输入与输出的最大长度总和，我们将此参数设定为“Infinite”以确保模型输出的完整性，避免因为超出最大令牌数而产生截断的情况。“Temperature”参数则负责调节输出结果的确定性，其取值在 0 到 1 之间。在此参数设定上，较低的“Temperature”值会使得模型生成的回答具有高度确定性和可预测性，即使多次以相同的提示向模型发起请求，也能获得高度一致或相同的答案；而较高的“Temperature”值则使得模型生成的回答更为多元化，但同时也可能增加结果偏离预期或主题的风险。基于上述考量，我们选择将“Temperature”设为 0.5，以平衡确定性和创新性。

ColBERT和MonoT5模型属于开源模型，可以下载到本地使用。ColBERT和MonoT5的模型参数大小分别为110M和220M。超参数的设定对模型性能有着直接影响，适当的调整超参数可以提升模型的准确度与效率。我们将ColBERT和MonoT5的BatchSize大小分别设置为 32 和 4 。

本文将RankGPT、ColBERT和MonoT5视为基准模型，并将与查询改写和文档得分融合机制结合后的模型视为它们相应的模型变体。

## 5 实验分析

### 5.1 整体性能分析

表 4 展示了在TREC-DL19、TREC-DL20以及TREC-COVID数据集上，不同检索模型在采用查询改写和文档得分融合机制前后的检索性能表现。其中，标有“w/o rewrite”的数据行展示了未进行查询改写和未采用文档得分融合机制时的检索性能；标有“w/ rewrite”的数据行展示了只采用查询改写不采用文档得分融合机制时的检索性能；标有“w/ fusion”的数据行展示了采用查询改写和文档得分融合机制后的检索性能；此外，n@k表示nDCG@k评价指标。为了直观地呈现本文提出的方法对检索性能的影响，表 4 仅列举了表现最佳的结果。至于不同长度的查询以及不同的文档得分融合情况对检索模型性能的影响将在 5.2 节和 5.3 节具体分析。

| Important Parameters | Parameter Value |
|---|---|
| OpenAI API | gpt-3.5-turbo |
| Max_tokens | Infinite |
| Temperature | 0.5 |

Table 3: gpt-3.5-turbo模型的重要参数设置

从表 4 中的实验数据得知，无论是通用领域的数据集还是专有领域的数据集，与查询改写和文档得分融合机制结合后的模型变体，其检索性能均优于相应的基准模型。


(c) 2024 中国中文信息学会计算语言学专业委员会

| | TREC-DL19 | | | TREC-DL20 | | | TREC-COVID | | |
|---|---|---|---|---|---|---|---|---|---|
| | n@1 | n@5 | n@10 | n@1 | n@5 | n@10 | n@1 | n@5 | n@10 |
| **BM25** | 50.77 | 49.02 | 47.95 | 56.48 | 49.65 | 49.36 | 75.00 | 67.51 | 62.32 |
| **ColBERT** | | | | | | | | | |
| w/o rewrite | 74.41 | 71.54 | 68.78 | 76.23 | 70.78 | 65.91 | 73.00 | 69.41 | 66.62 |
| w/ rewrite | **74.81** | 70.05 | 66.17 | 73.77 | 68.76 | 65.46 | 71.00 | **73.82** | **71.99** |
| w/ fusion | **79.85** | **74.01** | **71.50** | **79.32** | **72.87** | **69.08** | **76.00** | **75.22** | **72.62** |
| **MonoT5** | | | | | | | | | |
| w/o rewrite | 77.13 | 73.02 | 69.89 | 78.39 | 70.13 | 67.39 | 78.00 | 74.57 | 72.16 |
| w/ rewrite | 75.97 | 72.64 | 69.44 | 75.31 | 68.95 | 65.58 | 75.00 | **78.13** | **75.95** |
| w/ fusion | **79.06** | **76.19** | **71.68** | **80.86** | **72.84** | **68.34** | **81.00** | **79.93** | **78.00** |
| **RankGPT** | | | | | | | | | |
| w/o rewrite | 82.17 | 71.15 | 65.80 | 79.32 | 66.76 | 62.91 | - | - | 76.67 |
| w/ rewrite | 79.07 | 70.66 | **67.49** | 78.70 | **68.64** | **63.74** | - | - | **77.19** |

Table 4: 查询改写对不同检索模型的性能影响（n@k表示nDCG@k评价指标）

在通用领域数据集（DL19和DL20）上，我们发现当检索模型只采用查询改写不采用文档得分融合机制时，ColBERT和MonoT5模型的检索性能出现了下降，但RankGPT模型展现出了更好的检索性能。我们从不同检索模型架构的角度对这一现象进行了分析：(1) ColBERT是基于BERT的排序方法，主要关注查询与文档的精确匹配。改写后的查询虽更自然，但可能带来语义微调或额外信息，干扰原查询的匹配，导致检索性能下滑。(2) MonoT5是基于“Seq2Seq”的T5模型，预测查询与文档的匹配度。当改写后的查询偏离原意时，捕捉完整语义变的困难，从而导致检索性能减弱。(3) RankGPT使用ChatGPT理解查询的上下文和意图，并检索与之相关的文档。与前两种模型不同，RankGPT更能够适应和理解由ChatGPT改写的查询。这是因为改写后的查询往往更具有上下文信息和语言流畅性，这使得RankGPT能够更好地抓住查询的整体意图并进行有效的文档匹配。综上所述，对于更依赖于查询和文档之间精确匹配的模型（如ColBERT和MonoT5），ChatGPT的改写可能会降低它们的检索性能。但对于能够理解和适应自然语言查询的模型（如RankGPT），改写可能增强它们的检索能力，因为改写带来了更丰富的语境和语义信息。此外，这一实验结论也证明了文档得分融合机制的可行性和泛化性。

相比于通用领域数据集，在专业领域数据集（TREC-COVID）上，我们发现当检索模型只采用查询改写不采用文档得分融合机制时，ColBERT、MonoT5和RankGPT模型的检索性能都有显著提升。对于通用领域的数据集，我们认为一些查询包含的关键词可能拥有多重语义解释。因此，当ChatGPT对查询进行改写时，它们可能会基于和检索文档不相关的语义信息进行改写，导致了检索模型的性能下降。而TREC-COVID只关注新冠病毒的主题，这种单一和专业化的主题特点，使得ChatGPT在进行查询改写时，可以更为准确地捕捉到相关的语义信息，避免了因为语义的多样性而带来的改写误区。因此，我们推测对于具有明确主题和范围的数据集，ChatGPT在查询改写方面具有更好的表现，从而更有可能提升检索任务的性能。

### 5.2 不同长度查询的对比分析

表 5 中，“Len”一列记录了新查询的查询长度，$kx$ 表示新查询长度是原始长度的 $k$ 倍；n@k表示nDCG@k评价指标。根据表 5 的实验数据得知，当检索模型采用查询改写和文档得分融合机制时，不同的查询长度对检索模型的性能产生了显著的影响。从表 5 中可以看出，过于简短或冗长的查询都无法使得检索模型达到其最佳的性能水平。实验数据显示，一般情况下，查询长度达到原始长度的 5 倍或 7 倍时，检索模型的性能往往呈现出最佳表现。

这一实验结论可以从语义信息的丰富度来解释：对于长度较短的查询，其主要问题在于语义信息的匮乏。当查询词汇量有限时，其语义信息难以涵盖广泛的语义空间，导致检索模型难以准确把握用户的检索意图，也就增加了相关文档被遗漏的风险。相反，过度冗长的查询可能包含大量的非关键性信息，这不仅会增加信息处理的计算成本，还可能导致关键语义的丢失。此外，过长的查询往往导致文本中核心的语义信息被边缘化或模糊化，这使得检索模型难以识


(c) 2024 中国中文信息学会计算语言学专业委员会

| | | TREC-DL19 | | | TREC-DL20 | | | TREC-COVID | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Len | n@1 | n@5 | n@10 | n@1 | n@5 | n@10 | n@1 | n@5 | n@10 |
| **BM25** | - | 50.77 | 49.02 | 47.95 | 56.48 | 49.65 | 49.36 | 75.00 | 67.51 | 62.32 |
| **ColBERT** | | | | | | | | | | |
| w/o rewrite | - | 74.41 | 71.54 | 68.78 | 76.23 | 70.78 | 65.91 | 73.00 | 69.41 | 66.62 |
| | 1x | 75.19 | 71.58 | 69.43 | 77.78 | 71.11 | 67.68 | 77.00 | 72.48 | 71.56 |
| | 3x | 79.06 | 72.41 | 70.40 | **79.32** | **72.91** | 68.39 | 74.00 | 75.13 | 71.04 |
| w/ fusion | 5x | **79.85** | 73.74 | 70.63 | **79.32** | 72.87 | **69.08** | **78.00** | 75.06 | 71.28 |
| | 7x | **79.85** | **74.01** | **71.50** | 76.23 | 70.32 | 67.25 | 76.00 | **75.22** | **72.62** |
| | 9x | 74.41 | 72.40 | 70.42 | 79.01 | 71.66 | 69.06 | 73.00 | 70.91 | 68.60 |
| **MonoT5** | | | | | | | | | | |
| w/o rewrite | - | 77.13 | 73.02 | 69.89 | 78.39 | 70.13 | 67.39 | 78.00 | 74.57 | 72.16 |
| | 1x | 77.51 | 73.27 | 70.67 | **82.09** | **71.37** | **68.63** | 79.00 | 78.20 | 77.41 |
| | 3x | 78.29 | 74.32 | 71.36 | 80.86 | **72.84** | 68.34 | 76.00 | 78.80 | 76.97 |
| w/ fusion | 5x | 78.29 | 75.67 | **71.90** | 78.09 | 69.83 | 67.27 | **81.00** | **79.93** | **78.00** |
| | 7x | **79.06** | **76.19** | 71.68 | 76.23 | 69.46 | 66.16 | 77.00 | 75.55 | 73.10 |
| | 9x | 75.96 | 72.21 | 69.45 | 73.14 | 68.33 | 65.48 | 71.00 | 69.31 | 66.37 |

Table 5: 不同长度的查询对检索性能的影响（n@k表示nDCG@k评价指标）

别文本的重点，从而降低了检索模型在相关文档评分上的准确性。综上所述，为了优化检索模型的性能，合适的查询长度是至关重要的。在对查询进行改写优化时，关键在于找到一个平衡点，确保查询既不失去必要的语义信息，也不包含过多的冗余信息。

### 5.3 不同组合权重的对比分析

检索模型首先独立地评估与原始查询和新查询相关的每个文档的相关性。然后引入一个权重分配策略，该策略更倾向于给予得分较高的查询更高的权重，即认为这些查询与查询的真实意图更为接近，因此应在最终排名中占据更重要的位置。在进行文档得分融合时，我们会对原始查询的文档得分和新查询的文档得分分配不同的权重，即设置超参数 $\lambda$ 的值，如公式 1 所示。

图 3 展示了在进行文档得分融合时，不同的权重分配对检索性能的影响。当 $\lambda$ 值为 0 时，表示新查询的文档相关性得分；当 $\lambda$ 值为 1 时，表示原始查询的文档相关性得分。图 3(a) 和图 3(c) 分别展示了ColBERT模型在DL19数据集和DL20数据集上进行文档得分融合后的检索性能，在这两个通用领域数据集上，当原始查询的权重占比为 0.7 左右时，检索模型的性能达到最高。图 3(b) 和图 3(d) 分别展示了MonoT5模型在DL19数据集和DL20数据集上进行文档得分融合后的检索性能，同样当原始查询的权重占比为 0.7 左右时，检索模型的性能达到最高。图 3(e) 和图 3(f) 分别展示了ColBERT模型和MonoT5模型在Covid-19数据集上进行文档得分融合后的检索性能，在这个专有领域的数据集上，当原始查询的权重占比为 0.4 左右时，两个检索模型的性能达到最高。

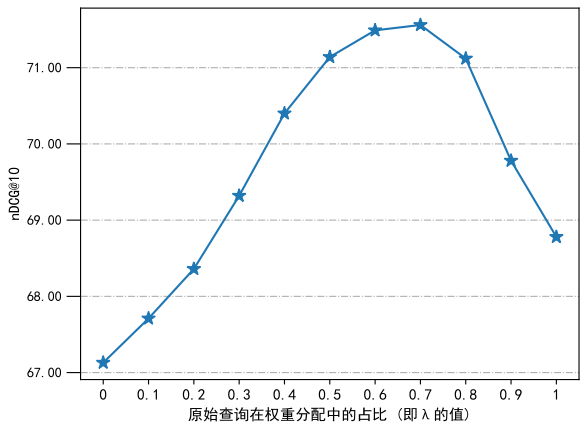
文档得分融合机制的实验结果表明，在通用领域的数据集中，超参数 $\lambda$ 会存在一定的泛化能力；而在专有领域（比如医学领域）的数据集中，超参数 $\lambda$ 也表现出一定的泛化能力。综上所述，超参数 $\lambda$ 在相同领域的数据集中存在一定的泛化能力。因此超参数 $\lambda$ 在不同领域的数据集中需要设置不同的值，但是在相同领域的数据集中可以设置相似的 $\lambda$ 值。另外，图 3 中的实验结果也证明了我们的权重分配策略是合理的，即更倾向于给予得分较高的查询更高的权重。

## 6 结语

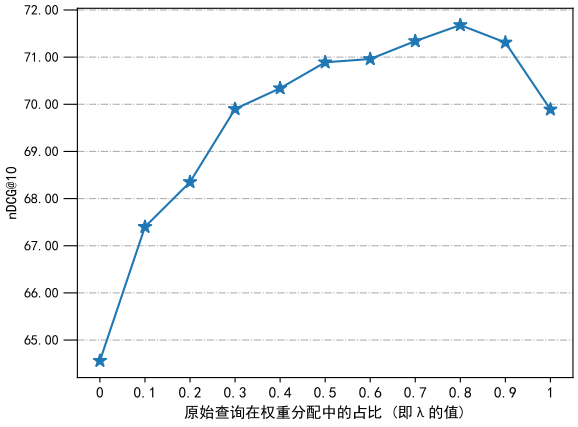
本文提出了一种基于ChatGPT查询改写的文档检索方法。面对大语言模型生成内容的不可控性，我们设计了一种文档得分融合机制，这一机制在实验中已被证明能够显著增强检索模型的性能。针对文本长度对于检索模型性能的影响，本文证明了存在一个最佳的查询长度区间，该区间内的查询文本能够带来最佳的检索性能。此外，在两个通用领域的数据集和一个专业领


(c) 2024 中国中文信息学会计算语言学专业委员会

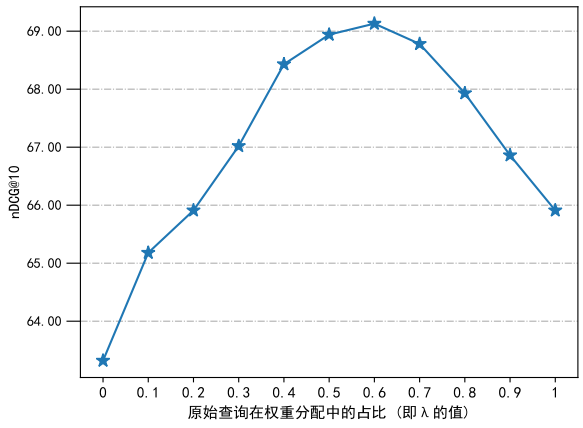
域的数据集上的大量实验表明，我们提出的方法有较好的可行性和泛化性。最后，我们未来的工作会考虑如何利用检索增强技术结合权威知识库的信息进一步完善和提高查询改写的效果。



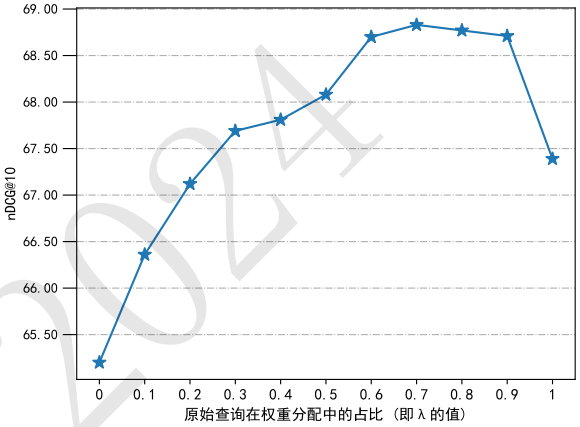

(a) ColBERT在DL19上进行文档得分融合



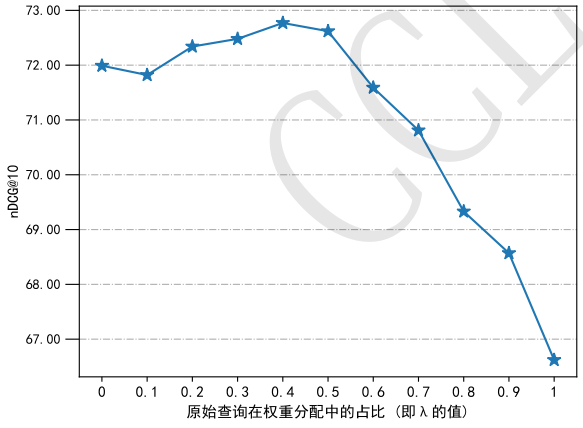

(b) MonoT5在DL19上进行文档得分融合



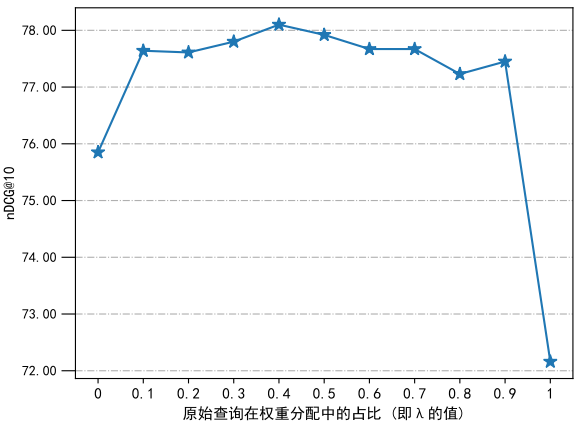

(c) ColBERT在DL20上进行文档得分融合




(d) MonoT5在DL20上进行文档得分融合




(e) ColBERT在Covid19上进行文档得分融合




(f) MonoT5在Covid19上进行文档得分融合

Figure 3: 文档得分融合时，不同的权重分配对检索性能的影响

## 参考文献

Nasreen Abdul-Jaleel, James Allan, W Bruce Croft, Fernando Diaz, Leah Larkey, Xiaoyan Li, Mark D Smucker, and Courtney Wade. 2004. Umass at trec 2004: Novelty and hard. *Computer Science Department Faculty Publication Series*, page 189.


(c) 2024 中国中文信息学会计算语言学专业委员会

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Gianni Amati and Cornelis Joost Van Rijsbergen. 2002. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems (TOIS)*, 20(4):357–389.

Hiteshwar Kumar Azad and Akshay Deepak. 2019. Query expansion techniques for information retrieval: a survey. *Information Processing & Management*, 56(5):1698–1735.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Guihong Cao, Jian-Yun Nie, Jianfeng Gao, and Stephen Robertson. 2008. Selecting good expansion terms for pseudo-relevance feedback. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 243–250.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M. Voorhees. 2020. Overview of the TREC 2019 deep learning track. *CoRR*, abs/2003.07820.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. 2021. Overview of the TREC 2020 deep learning track. *CoRR*, abs/2102.07662.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Rolf Jagerman, Honglei Zhuang, Zhen Qin, Xuanhui Wang, and Michael Bendersky. 2023. Query expansion by prompting large language models. *arXiv preprint arXiv:2305.03653*.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48.

Hang Li, Shengyao Zhuang, Ahmed Mourad, Xueguang Ma, Jimmy Lin, and Guido Zuccon. 2022. Improving query representations for dense retrieval with pseudo relevance feedback: A reproducibility study. In *European Conference on Information Retrieval*, pages 599–612. Springer.

Hang Li. 2022. *Learning to rank for information retrieval and natural language processing*. Springer Nature.

Craig Macdonald, Nicola Tonellotto, Sean MacAvaney, and Iadh Ounis. 2021. Pyterrier: Declarative experimentation in python from bm25 to dense retrieval. In *Proceedings of the 30th acm international conference on information & knowledge management*, pages 4526–4533.

Iain Mackie, Shubham Chatterjee, and Jeffrey Dalton. 2023. Generative relevance feedback with large language models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2026–2031.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Ronak Pradeep, Rodrigo Nogueira, and Jimmy Lin. 2021. The expando-mono-duo design pattern for text ranking with pretrained sequence-to-sequence models. *arXiv preprint arXiv:2101.05667*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

CCL 2024


(c) 2024 中国中文信息学会计算语言学专业委员会

Weiwei Sun, Lingyong Yan, Xinyu Ma, Pengjie Ren, Dawei Yin, and Zhaochun Ren. 2023. Is chatgpt good at search? investigating large language models as re-ranking agent. *arXiv preprint arXiv:2304.09542*.

Yi Tay, Mostafa Dehghani, Vinh Q Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Siamak Shakeri, Dara Bahri, Tal Schuster, et al. 2022. Ul2: Unifying language learning paradigms. *arXiv preprint arXiv:2205.05131*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. 2021. Trec-covid: constructing a pandemic information retrieval test collection. In *ACM SIGIR Forum*, volume 54, pages 1–12. ACM New York, NY, USA.

Liang Wang, Nan Yang, and Furu Wei. 2023a. Query2doc: Query expansion with large language models. *arXiv preprint arXiv:2303.07678*.

Xiao Wang, Craig Macdonald, Nicola Tonellotto, and Iadh Ounis. 2023b. Colbert-prf: Semantic pseudo-relevance feedback for dense passage and document retrieval. *ACM Transactions on the Web*, 17(1):1–39.

Tianyu Wu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, Qing-Long Han, and Yang Tang. 2023. A brief overview of chatgpt: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, 10(5):1122–1136.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Zhi Zheng, Kai Hui, Ben He, Xianpei Han, Le Sun, and Andrew Yates. 2020. Bert-qe: contextualized query expansion for document re-ranking. *arXiv preprint arXiv:2009.07258*.

CCL 2024

# 基于汉语字词资源的检索增强生成与应用评估

**殷雅琦[1,2]，刘扬[1,2]*，王悦[1,2]，梁启亮[1,3]**
[1]北京大学多媒体信息处理全国重点实验室，北京 100871
[2]北京大学计算机学院，北京 100871
[3]北京大学信息科学技术学院，北京 100871
yyqi@stu.pku.edu.cn，{liuyang，wyy209}@pku.edu.cn，lql_eecs@qq.com

## 摘要

汉语遵循“由字组词，由词造句”的原则，字词相关信息是一类基础且关键的计算资源。在大语言模型时代，挖掘并评价该类资源的效用是增强模型语言能力的一个重要研究方面。作为有效促进资源与模型结合的一种方式，检索增强生成目前在该类资源上的应用大都关注模型未学习过的濒危语言，其在模型已学习过语言上的潜在价值有待挖掘。本文基于语言学的视角，构建具有良好例句覆盖率与丰富度的字词资源，并借助检索增强生成技术路线，探索这类资源与不同任务、模型的结合方法。评估实验表明，该方法在所有实验模型与任务中均带来了显著的准确率提升，平均达4.78%，其中，在语素义消歧、词义消歧与隐喻识别任务中分别提升了6.91%、4.24%和3.19%，这展示出字词资源对模型的语言准确理解能力的潜在价值。这些资源构造、方法探索和应用评估，为语言学资源与大语言模型的结合提供了新的思路与方法。

**关键词：** 汉语字词；资源建设；检索增强生成；应用评估

# Chinese Character- and Word-Based Retrieval Augmented Generation and Application

**Yaqi Yin[1,2], Yang Liu[1,2*], Yue Wang[1,2], Qiliang Liang[1,3]**
[1]National Key Laboratory for Multimedia Information Processing,
Peking University, Beijing 100871
[2]School of Computer Science, Peking University, Beijing 100871
[3]School of Electronics Engineering and Computer Science,
Peking University, Beijing 100871
yyqi@stu.pku.edu.cn，{liuyang，wyy209}@pku.edu.cn，lql_eecs@qq.com

## Abstract

The Chinese language's 'character, word, sentence' structure makes character and word-related information essential resources. In the era of Large Language Models (LLMs), exploring and evaluating these resources is crucial for enhancing the language capabilities of models. Retrieval Augmented Generation (RAG) is an effective means of integrating resources with models, currently focusing on languages not learned by the model, leaving untapped potential in learned languages. In this study, with a linguistic perspective, we construct a resource with improved contextual coverage and utilize RAG to combine knowledge with various tasks and models. Experimental results show accuracy improvements across all tested models and tasks with 4.78% on average.

*通讯作者
**基金项目**：国家自然基金项目（62036001）、国家社科基金项目（18ZDA295）
©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

Specifically, it has achieved increases of 6.91%, 4.24% and 3.19% in morpheme sense disambiguation, word sense disambiguation, and metaphor recognition. The results show the potential of these resources for enhancing language understanding capabilities. These resource constructions, method explorations, and application evaluations offer new insights for integrating linguistic knowledge bases with LLMs.

**Keywords:** Chinese Character and Word , Resource Construction , Retrieval Augmented Generation , Application and Evaluation

## 1 引言

汉语中汉字是最小的独立表义单位，向上遵循“由字组词，由词造句”的原则(朱德熙, 1982)。这意味着字及其相关信息是认知词义的关键，而词及其相关信息又是理解句义的关键。因此，对汉语字词进行语义等信息标注的资源，作为文本内容理解的媒介与载体，已成为自然语言处理中不可或缺的基础和关键的计算资源(吉志薇and 冯敏萱, 2015; 刘扬et al., 2018)。利用这些资源，相关研究在词义消歧(Luo et al., 2018; Zheng et al., 2021b)、释义生成(Yang et al., 2020; Zheng et al., 2021a)、隐喻识别(Su et al., 2020)等任务中显著提高了模型的效果。

近年来，以ChatGPT[1]为代表的大型生成式语言模型(Touvron et al., 2023; Bai et al., 2023)由于其优越的多语言能力(Zhao et al., 2023; Zhang et al., 2024c)和任务表现(Zhong et al., 2023; Huang et al., 2023)，引起了行业的广泛关注。然而它们在对语言的准确理解方面仍存在不足(Martínez et al., 2023)，尤其是非英语语言(Robinson et al., 2023; Bang et al., 2023)，例如中文(Hendy et al., 2023; Ahuja et al., 2023; Liu et al., 2023)。因此，挖掘并评价字词资源在增强模型语言能力方面的效用并加以应用评估，是一个重要的研究方面。

在此研究方面，检索增强生成（Retrieval Augmented Generation, RAG）(Lewis et al., 2020)是一种被广泛研究与采纳的形式。它利用检索系统从外部资源中获取的相关信息，增强模型生成内容的质量与准确度。许多研究表明，相较于微调，这一方法不仅支持持续、低成本的知识更新和特定领域信息的集成(Strubell et al., 2019; Balaguer et al., 2024)，还能更有效、稳定地提高模型的准确度和可信度(Ovadia et al., 2024; Balaguer et al., 2024)，且可以应用于黑盒模型(Zhang et al., 2024b)，已逐渐成为字词资源在大语言模型中的重要应用方法。这些已有的基于字词资源的检索增强生成方法(Tanzer et al., 2023; Zhang et al., 2024a)利用字词典与语法手册，提高模型在机器翻译、对话理解等任务上的表现。目前，此类研究大都集中于模型未学习过的濒危语言，它们在模型已学习过的语言上的潜在价值尚有待挖掘。

我们注意到，对于非英语语言，即使是已针对该语言进行训练的模型，这类资源与方法也能为其提供有效的知识，优化其在下游任务中的表现。如图1所示，在面向不同语言单位的汉语消歧任务中，字词释义信息的注入帮助GPT-3.5模型[2]提供了正确的预测结果。

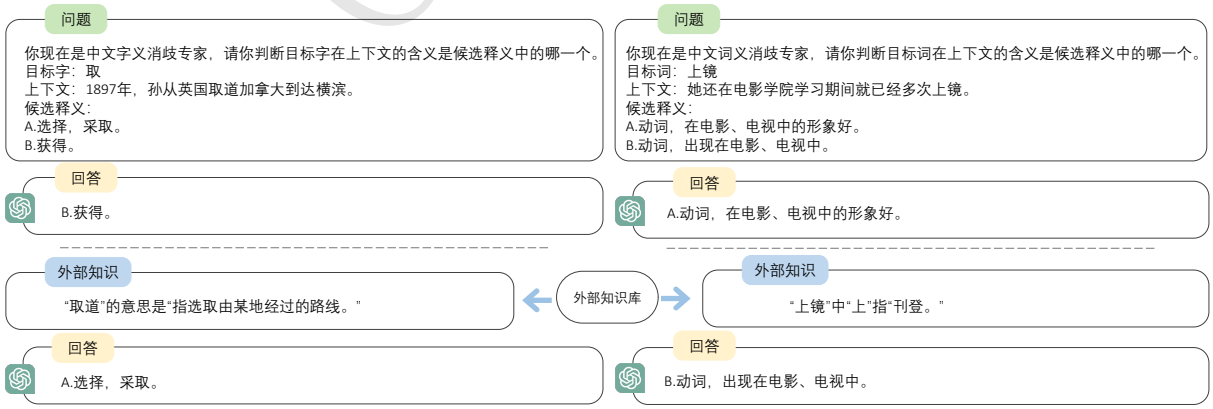



Figure 1: 基于外部知识的模型语言理解能力增强例子

[1] https://platform.openai.com/docs/models/gpt-3-5
[2] 本文采用的均为GPT-3.5-turbo-0613，以下简称为GPT-3.5。


(c) 2024 中国中文信息学会计算语言学专业委员会

受限于资源建设的不足和有效数据的匮乏，前期工作大多忽略了这方面的考察。本文希望从语言学角度出发，探索字词资源对大语言模型的汉语准确理解能力的效度：期望解决的一个问题是基于已有的受限资源，如何有效开展面向检索增强生成的知识工程，以更好地满足相关计算需求；在此基础上，另一个问题是如何实现知识工程获得的批量信息与不同模型的有效结合，并验证其有效性。

## 2 相关工作

### 2.1 汉语字词信息相关知识工程

汉语遵循从字、词、短语到句子的层级结构，因此，对字与词进行语义标注的资源是自然语言处理领域中重要的计算资源。在相关的资源建设上，语素作为汉语中最小的音义结合体(朱德熙, 1982)，在构词中非常活跃，处于重要的地位(尹斌庸, 1984; 徐枢, 1990)，语素义的组合在一定程度上能体现词义(符淮青, 1981)。同时，语素和语素义具有强大的表义能力(Zhang, 1997; Luo, 2013)，因此数量相对稳定，有助于相关资源的构建。

在相关资源建设上，**“汉语语素数据库”**(苑春法and 黄昌宁, 1998) 覆盖了6,763个常用汉字的语素项信息，并对其构成的二字复合词进行了结构描述和意义绑定。该数据库中各义项之间的连接欠缺，限制了其在计算中的深入应用；**“汉字义类信息库”**与**“汉语语义构词信息库”**(亢世勇et al., 2004)包含了6,763个常见汉字的字位及52,366个二字词的义类与释义，并与《同义词词林》(梅家驹et al., 1983)进行语义绑定。其归类以现有《词林》为标准，存在义项的因果参照问题；**“词素-义类数据库”**(吉志薇and 冯敏萱, 2015)提取了《现代汉语词典》（以下简称《现汉》）中2,268个词素并标注其义类，基于此整理了8,984个二字词的词素意义和词素间的词化意义。其收录的二字词采样不均衡且数据规模过小，难以满足大规模计算需求；**“汉语概念词典”**(刘扬et al., 2018)提取和编码了《现汉》中8,514个汉字的20,855个语素义，并构建语义类别，在此基础上，对《现汉》中41,474个二字词的52,108个义项分别进行语素义绑定。这一资源覆盖了汉语中常用字与罕见字的情况，有助于大规模计算与准确理解应用的开展。

这些前期工作，为基于语言学的计算路径提供了数据基础，并使基于汉语字词的检索增强生成成为可能。

### 2.2 大语言模型检索增强生成方法

检索增强生成技术(Lewis et al., 2020; Guu et al., 2020)旨在通过对外部资源的动态信息检索，为语言模型提供额外的信息。随着具有强大上下文学习能力的大语言模型的出现，这类方法倾向于在现有模型之上补充性地集成检索功能(Ram et al., 2023; Shi et al., 2023)，将原始输入序列与检索到的文本拼接起来作为输入，在问答(Guo et al., 2024; Zhang et al., 2024b)、对话(Peng et al., 2023; Kang et al., 2023)、个性化文本生成(Mysore et al., 2023)等自然语言处理任务中显著提升了基础模型的表现，在代码生成(Sun et al., 2023)、医疗(Chen et al., 2024)、金融(Kang and Liu, 2023)、法律(Savelka et al., 2023)等领域下也有广泛的应用。

在这些面向实体知识或领域知识的检索增强生成之外，部分工作注意到语言知识的潜在价值，尤其是对于非英语的低资源语言。相关工作按知识的来源可以分为两类：第一类采用自动生成的语言知识，如Lu(2023)利用NLLB机器翻译模型(Team et al., 2022)对关键词进行多语言翻译，Gao(2023)利用Stanza(Qi et al., 2020)生成词性标注标签序列。通过将自动生成的语言知识整合到提示词中，两种方法均在多种语言上提高了模型的机器翻译效果。但它们的语言知识质量与范围受知识生成模型的能力限制，在文化相关的或长尾的知识方面存在挑战；第二类则采用权威的语言学知识库，如Tanzer(2023)使用词典和语法手册辅助大型语言模型翻译预训练阶段未接触过的语言。Zhang(2024a)则利用语素知识库和语法手册，在多种下游任务中增强了模型在濒危语言上的表现。这些工作有效地将语言学资源引入大语言模型的检索增强生成，但大都关注模型未学习过的濒危语言，忽略了这些资源对于更多语言的潜在价值。

前人的工作证实了检索增强生成的有效性，为汉语字词资源与大语言模型的结合提供了可行的技术路径。

## 3 面向检索增强生成的汉语字词知识工程

为了更好地利用字词资源提升大语言模型的汉语准确理解能力，本文基于前人提出的知识库，开展面向检索增强生成的知识工程。


(c) 2024 中国中文信息学会计算语言学专业委员会

### 3.1 数据来源基础

在满足计算需求的基础上，检索增强生成所需的资源需要对语言具有良好的覆盖度。本文采用“汉语概念词典”作为基础资源，其中语素作为字的不同使用及意义表征。该资源内的字词信息如表1所示。这一资源发掘了《现汉》中8,514个字的20,855个义项，其中12,783个义项包含了27,215条例句，占全部的61.29% 。该资源同时包含53,780个二字及多字词的65,335个义项，其中29,669个义项包含了42,163条例句，占全部的45.41%。

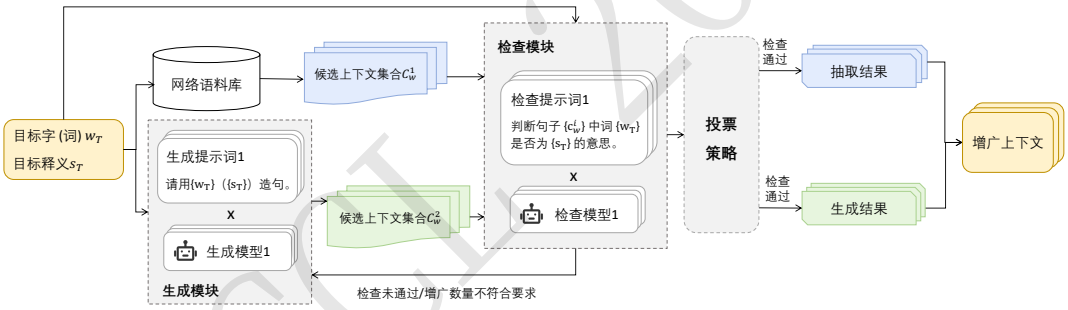
| **语素** | 取 | **词** | 上镜 |
|---|---|---|---|
| **语素义** | 选择；采取。 | **词义** | 出现在电影、电视中。 |
| **语素类** | 动语素 | **词类** | 动词 |
| **例句** | 1897年，孙从英国取道加拿大到达横滨。 | **例句** | 她还在电影学院学习期间就已经多次上镜。 |

Table 1: “汉语概念词典”中的字词知识示例

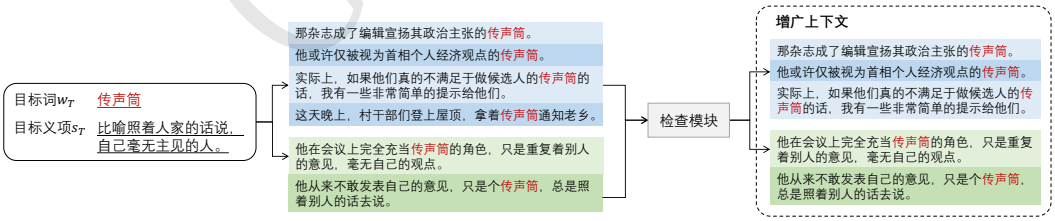
基于“汉语概念词典”，FiCLS数据集(Zheng et al., 2021b)对其中多义且有例句的部分进行了例句文本长度和数量的扩充。具体来说，FiCLS数据集包含2,706个字的10,639个义项及92,135条例句，平均每个义项8.66条。该数据集同时包含4,358个二字词的9,597个义项及27,061条例句，平均每个义项2.82条。

迄今为止，这些资源尚无法为《现汉》中的全部义项提供足够的例句信息，限制了它们在下游任务上的充分应用。本文提出利用网络语料库与生成式语言模型，在“汉语概念词典”与FiCLS数据集的基础上，对其中无例句或少例句的义项进行增广，为后续的实验与评估提供数据基础。

### 3.2 增广方法与设置




(a) 增广流程




(b) 增广示例

Figure 2: 面向检索增强生成的汉语字词知识工程

本文提出的增广方法如图2a所示。对于目标字（词）$w_T$的释义$s_T$，我们通过两种方法，提高增广例句的可靠性与多样性，同时保障增广数量：第一种方法基于网络语料库，从中抽取包含$w_T$的候选例句集合$C_w^1$；第二种方法基于生成式模型，结合多种模型与提示词为$w_T$生成候选例句集合$C_w^2$。所有候选例句均会通过检查模块，最终由投票策略判断是否通过。图2b中给出了词“传声筒” 的增广示例，蓝色与绿色部分分别为来自两种方法的例句。可以看出，检查模块有效地排除了不符合$s_T$的增广结果；同时，网络语料库中的文本比自动生成的更为灵活多样。

在**网络语料库构建**方面，考虑到增广例句的多样性，我们选择人民日报[3]、造句网[4]与古诗词网[5]作为语料来源。其中，人民日报文本数量充足，但表述相对正式、单一；造句网文本来源广泛、表述灵活，但数量有限；古诗词网则囊括大量古代诗词典籍，能有效增广部分在古汉语中更常见的字词。语料库最终涵盖人民日报2018-2020语料2,573,625句，包含69,032,189词；造句网语料447,350句，包含10,448,806词；古诗词网语料55,628句，包含8,463,423词。

在**生成模块**与**检查模块**中，我们采用三个在多种评测中表现良好的大语言模型：GPT-3.5、Qwen-max(Bai et al., 2023)与ChatGLM-turbo(Zeng et al., 2022; Du et al., 2022)，后两个为面向中文的模型，保证了良好的中文理解与生成能力。模型参数保留为默认参数，以获得最佳效果。模型提示词通过小批量验证的方式选取（附录A.1）。生成模块中，3个模型轮流用提示词生成例句；检查模块中，3个模型用所有提示词产生共9条投票结果。

依据上述投票结果，由**投票策略**决定是否保留该候选例句。其中，阈值$\#neg \leq 0$表示仅反对票数量$\leq 0$的例句会被认为检查通过。为了探究不同阈值对投票结果的影响，我们从各来源的例句中分别抽取300条，人工评估其正确性。结果如表2所示。其中，$P$为准确率，$R$为召回率，“单义”代表《现汉》中只有一条义项的字或词。最严格的投票策略（$\#neg \leq 0$）带来的准确率最高，但召回率极低，尤其是在“古诗词网”语料中。这主要是由于模型对古汉语的理解较差，因此倾向于在检查模块中给出“否”的判断结果。为了在保证质量的前提下尽量扩大增广数量，本文设置投票策略如下：1. 单义字词的阈值为$\#neg \leq 3$、古诗词网语料的阈值为$\#neg \leq 4$；2. 对于不属于1中的其余候选例句，先按阈值$\#neg \leq 0$进行第一轮筛选。对于筛选后未得到增广的义项，再按阈值$\#neg \leq 1$补召回。

| | 阈值 | 基于语料库 | | | | | | 基于生成 | |
|---|---|---|---|---|---|---|---|---|---|
| | | 人民日报 | | 造句网 | | 古诗词网 | | | |
| | | $P$ | $R$ | $P$ | $R$ | $P$ | $R$ | $P$ | $R$ |
| **全集** | $\leq 0$ | **95.08** | 61.70 | **96.77** | 60.61 | **100.00** | 16.06 | **96.88** | 24.80 |
| | $\leq 1$ | 94.44 | 72.34 | 95.71 | 67.68 | 97.50 | 28.47 | 94.34 | 40.00 |
| | $\leq 2$ | 91.76 | 82.98 | 93.83 | 76.77 | 98.44 | 45.99 | 95.77 | 54.40 |
| | $\leq 3$ | 88.66 | 91.49 | 90.22 | 83.84 | 98.88 | 64.23 | 91.92 | 72.80 |
| | $\leq 4$ | 88.00 | **93.62** | 87.13 | **88.89** | 98.15 | **77.37** | 86.61 | **77.60** |
| **单义** | $\leq 0$ | 97.44 | 69.09 | **100.00** | 57.50 | **100.00** | 16.06 | 94.12 | 15.69 |
| | $\leq 1$ | 97.83 | 81.82 | 96.30 | 65.00 | 97.50 | 28.47 | 96.00 | 23.53 |
| | $\leq 2$ | **98.00** | 89.91 | 96.97 | 80.00 | 98.44 | 45.99 | **97.06** | 32.35 |
| | $\leq 3$ | 96.30 | **94.55** | 97.22 | 87.50 | 98.88 | 64.23 | 92.86 | **38.24** |
| | $\leq 4$ | 94.55 | **94.55** | 95.00 | **95.00** | 98.15 | **77.37** | 82.98 | **38.24** |

Table 2: 例句增广中，检查模块在不同阈值下的准确率与召回率

在**实验设置**方面，由于网络语料库数据量过大，在基于语料库的增广过程中，对于拥有$K$个不同义项的字或词，我们为每个义项随机抽取$10 \times K$个候选例句。考虑到网络语料库的例句质量更高，我们优先采用基于网络语料库的方法，增广数量不足3条的用基于生成式模型的方法进行补充。

### 3.3 增广结果与分析

通过基于语料库的方法，我们共得到增广例句498,126条，其中，342,867条来自人民日报语料，149,390条来自造句网语料，5,869条来自古诗词网语料。在基于语料库的方法后，仍有17,876条语素义、31,239条词义的增广例句数量不足3条。通过基于生成式模型的方法，我们进一步获得增广例句28,724条。最终，本文得到增广例句526,850条。为确保其可靠性，我们分别从字与词的增广例句中随机抽取1,000条进行人工评测，统计与评测结果如表3所示。

我们将前人构建的资源（“汉语概念词典”与FiCLS）中的数据合并去重后得到187,642条增广前例句，与本文的增广例句合并后共得到714,492条全集例句，统计信息如表4所示。其中，“覆盖率(%)”为包含例句的义项比例。从统计结果可以看出，增广后例句的平均数量和覆盖率都有了显著提高，覆盖率均达到80%以上，展示出这一增广的显著效果。

[3] http://paper.people.com.cn/
[4] https://zaojv.com/
[5] https://www.gushici.com/


(c) 2024 中国中文信息学会计算语言学专业委员会

| | 新增例句数 | 准确率(%) | 平均例句长度 | 字词数 | 义项数 |
|---|---|---|---|---|---|
| 语素 | 33,022 | 96.00 | 39.90 | 7,182 | 10,701 |
| 词 | 493,828 | 99.00 | 47.84 | 46,158 | 53,752 |
| 全集 | 526,850 | 97.50 | 47.34 | 53,340 | 64,453 |

Table 3: 增广例句统计结果与人工评测

| 类型 | 字 | | | | 词 | | | |
|---|---|---|---|---|---|---|---|---|
| | 例句数 | 平均例句数 | 平均例句长度 | 覆盖率(%) | 例句数 | 平均例句数 | 平均例句长度 | 覆盖率(%) |
| 增广前 | 118,625 | 5.69 | 3.66 | 61.49 | 69,017 | 1.06 | 7.31 | 45.90 |
| 增广后 | 151,647 | 7.27 | 23.43 | 83.54 | 562,845 | 8.61 | 43.70 | 89.43 |

Table 4: 例句增广前后统计指标对比分析

## 4 基于汉语字词资源的检索增强生成方法

考虑到方法的适用性，本文利用检索增强生成，探索字词资源在大语言模型的汉语准确理解能力上的效度，如图3所示，包含基于关键词的知识检索与基于评估的知识排序两个流程。其中，我们以第3节得到的字词资源用作外部资源。该方法的输入信息包含用于获取关键词的文本$C_{keyword}$，以及用于计算相似度的文本$C_{sim}$，二者均来自任务的原始输入信息（本节的主要符号对照表见附录C）。

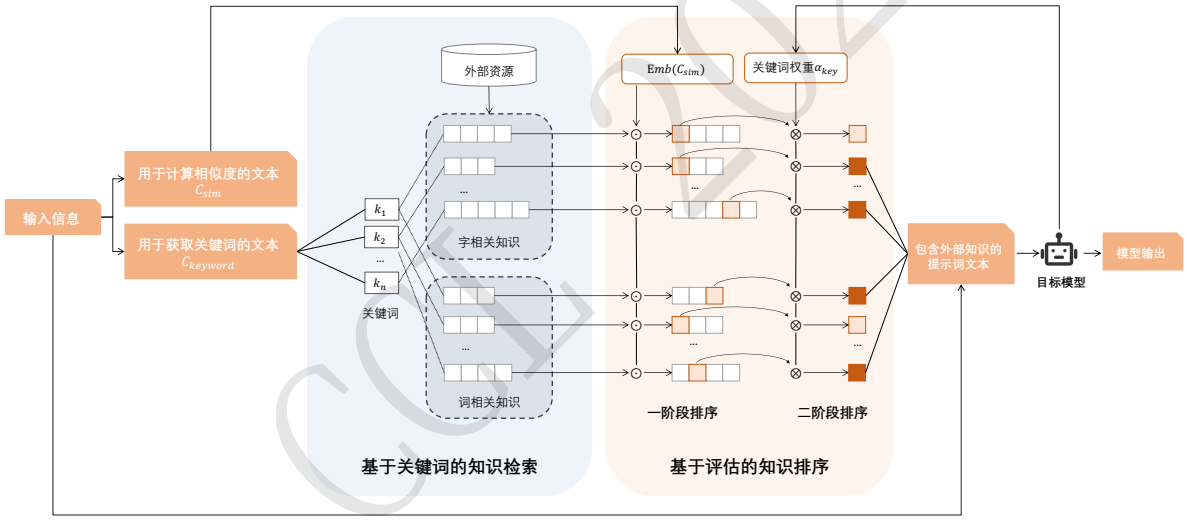



Figure 3: 基于汉语字词资源的模型增强方法

### 4.1 流程一：基于关键词的知识检索

基于关键词的知识检索负责基于目标文本，从字词资源中获取可能的相关知识。知识检索的目的是将输入文本与知识进行有效匹配，减少无关信息的引入，这一流程包含以下两步：

第一步，从用于获取关键词的文本$C_{keyword}$中抽取关键词集合$K = \{k_1, k_2, ..., k_n\}$，将文本转为满足字词资源需求的形式。这一流程旨在减少无关信息的引入，提高知识注入的效率；

第二步，利用关键词$k_i$检索字词资源，得到候选知识集合$D_i = \{d_{i,1}, d_{i,2}, ..., d_{i,l}\}$，其中包含释义信息，以及例句中的释义信息。表5展示了字“取”和词“取道”作为关键词时，分别获取到的部分知识示例。

### 4.2 流程二：基于评估的知识排序

基于评估的知识排序负责基于已有的候选知识，筛选得到最终注入到提示词中的部分。知

| | | |
|---|---|---|
| **字相关知识** | 释义信息 | “取”的意思是“选择，采取”。 |
| | 例句中的释义信息 | “取”在“取道”中的意思是“选择，采取”。 |
| **词相关知识** | 释义信息 | “取道”的意思是“指选取由某地经过的路线”。 |
| | 例句中的释义信息 | “取道”在“1897年，孙从英国取道加拿大到达横滨。”中的意思是“指选取由某地经过的路线”。 |

Table 5: 从字词资源中获取到的字词相关知识示例

识注入的目的是通过提供与目标任务相关的、模型理解不稳定或有困难的知识，提高模型对任务的能力。因此，准确筛选候选知识，关键在知识与输入信息的相关性，以及知识对目标模型的有效性。基于以上考虑，这一流程分为以下两个阶段：

第一阶段排序负责从关键词下的所有知识中，筛选得到与用于计算相似度的文本$C_{sim}$最相关的$N_{stage1}$个。考虑到候选知识除释义之外还包含例句，且总体目的与传统的词义消歧任务不完全一致，本节采用句向量之间的余弦相似度作为一阶段排序分数。对于关键词$k_i$的第$j$条候选知识$d_{i,j}$，它与$C_{sim}$的余弦相似度计算方式为：

$$cos\_similarity = cos(Emb(d_{i,j}), Emb(C_{sim})) = \frac{Emb(d_{i,j}) \cdot Emb(C_{sim})}{|Emb(d_{i,j})| \times |Emb(C_{sim})|} \in [-1, 1] \quad (1)$$

其中，$Emb$表示获取文本的句嵌入向量。考虑到本文只需要知识之间的相对顺序而非绝对数值，为了便于后续计算，我们定义第一阶段的排序分数如下：

$$score_{stage1} = 1 + cos\_similarity \in [0, 2] \quad (2)$$

在此基础上，第二阶段排序则负责从前一阶段得到的所有知识中，筛选出对模型最有效的$N_{stage2}$个。具体来说，我们希望获取目标模型理解最困难的知识，适当增加其权重，提高它们被注入到提示词中的概率，进而提高注入知识对目标模型的有效性。本文采用了此前MorphEval(Yin et al., 2024)中的评估结果。该工作在“汉语概念词典”的基础上，构建了基于汉语语素的大语言模型评估数据集，用于评估模型的语言准确理解能力。对于候选知识$d_{i,j}$，我们获取其在MorphEval数据集中出现的总次数$T_{i,j}$，以及模型对其预测错误的次数$E_{i,j}$。利用这两项数据，我们计算其权重$\alpha_{i,j}$如下：

$$\alpha_{i,j} = e^{\frac{E_{i,j}}{T_{i,j}}} \in [0, e] \quad (3)$$

在此基础之上，我们定义第二阶段的排序分数为：

$$score_{stage2} = \alpha \times score_{stage1} \in [0, 2e] \quad (4)$$

最终，我们利用为不同任务与不同类型知识构造的提示词模版，将获取到的$N_{stage2}$条知识融入到原提示词中，作为当前任务的输入信息。

## 5 检索增强生成的应用评估

### 5.1 语言准确理解评估任务的选取

为了验证汉语字词资源对大语言模型的汉语准确理解的有效性，我们选择了多个与语言的准确理解紧密相关的词内、词间自然语言处理任务进行评估：

**一、汉语语素义消歧任务（Morpheme Sense Disambiguation, MSD）**(Wang et al., 2024)旨在预测目标语素在特定上下文中的含义。该任务的消歧对象包含自由语素与黏着语素（详细介绍见附录B）。此前的相关工作基于BERT模型(Wang et al., 2024)，我们在本文将这一任务引入到对自回归语言模型的评测中，为其构建适合模型的选择题任务形式：给定语素、语素所在的上下文和所有语素义，要求模型选择上下文中的语素属于哪一个语素义。

**二、汉语词义消歧任务（Word Sense Disambiguation, WSD）**(Navigli, 2009)旨在预测目标词在特定上下文中的含义，是最核心的词级自然语言处理任务之一。该任务的消歧对象


(c) 2024 中国中文信息学会计算语言学专业委员会

除可独立成词的自由语素之外，还包含二字及多字词。此前的相关工作包括两种常见方法：基于知识的方法利用现有的大规模知识资源来推断上下文中的词义(Galley and McKeown, 2003; Agirre et al., 2014)，基于监督学习的方法则利用带有义项标注的数据集训练分类器(Huang et al., 2019; Scarlini et al., 2020; Zheng et al., 2021b)。然而这些方法采用的都是BERT式自编码模型，忽略了自回归模型的评估与应用，尤其是诸如GPT-3.5的大型生成式语言模型。本文将这一任务引入到对大语言模型的计算中，并沿用之前的实验设计，为这类模型构建选择题的任务形式。

**三、汉语隐喻识别任务（Metaphor Detection, MD）**旨在判断与识别文本中字面义和非字面义处于不同的语义域中的现象（详细介绍见附录B）。隐喻是蕴含丰富文化背景的语言现象，普遍存在于人类日常交流中(Richards, 1936)，因此对大语言模型至关重要。此前的相关工作同样大多基于BERT式自编码模型(Su et al., 2020; Choi et al., 2021; Li et al., 2023)，对自回归模型的关注较少，其中，Wachowiak等(2023)在英语与西班牙语中利用RiC（Reasoning in Conversation）提高GPT-3的隐喻源域预测能力，Tong等(2024)将任务转换为改写形式，在英语上评估多种生成式模型的表现。考虑到本文主要探究字词信息的知识注入，我们选择词级隐喻识别任务作为评估，即判断上下文中的目标词是否包含隐喻现象。我们参考前人方法(Comșa et al., 2022; Zhou et al., 2023; Tong et al., 2024)，为模型构建了选择题、判断题与填空题三种任务形式。

## 5.2 实验数据及其统计信息

**语素义消歧任务**中，为了获取能完整覆盖全部义项的评估数据集，本文沿用Wang等(2024)中提出的方法，从“汉语概念词典”中抽取语素、词与上下文信息，保留其中多义语素的部分。数据集共23,914条，根据上下文信息的不同可以分为词内和句内语素义消歧数据集。

**词义消歧任务**中，考虑到输入特征与语素义消歧任务的一致性，本文采用FiCLS数据集(Zheng et al., 2021b)用于评估。由于其全集数据的人工评测准确度有限，我们仅采用其中来自原始语料库而非自动增广的部分。数据集共28,770条，按消歧对象的不同可以分为单字词、二字词与多字词数据集。

**词级隐喻识别任务**中，本文采用陈龙等(2019)提出的非字面义数据集作为评估。这一数据集发掘了《现汉》中的3,524个非字面义词，包含隐喻和转喻两种词级现象，对应4,301个数据条目，与语素义消歧与词义消歧任务中的语素、语素义和词义等信息保持了一致性。

三个任务的数据集统计信息如表6所示。其中隐喻识别任务数据无释义，因此不做统计。

| 任务 | 数据集 | 数据条目数 | 字词数 | 义项数 | 上下文数 | 平均上下文长度 |
|---|---|---|---|---|---|---|
| 语素义消歧 | 词内消歧 | 13,276 | 5,519 | 13,276 | 12,356 | 2.15 |
| | 句内消歧 | 10,638 | 2,706 | 10,638 | 10,591 | 42.01 |
| 词义消歧 | 单字词 | 7,685 | 2,133 | 6,515 | 6,506 | 34.10 |
| | 二字词 | 19,316 | 8,683 | 19,316 | 19,239 | 27.45 |
| | 多字词 | 1,769 | 828 | 1,746 | 1,763 | 23.25 |
| 隐喻识别 | 选择形式 | 3,313 | 3,705 | - | 5,922 | 8.57 |
| | 判断形式 | 5,988 | 3,730 | - | 5,975 | 8.46 |
| | 填空形式 | 5,988 | 3,730 | - | 5,975 | 8.46 |

Table 6: 实验数据统计信息

## 5.3 实验方法与设置

在**实验设置**方面，在基于关键词的知识检索中，我们设置$C_{keyword}$与$C_{sim}$为上下文，并采用pkuseg分词器(Luo et al., 2019)将$C_{keyword}$切分为字词。在基于评估的知识排序中，对于两阶段检索的数量，为保证提示词总长度在模型预训练长度范围之内，我们设置$N_{stage1} = 3$，$N_{stage2} = 3$，即在一阶段排序后，每个关键词最多保留3条候选知识；在二阶段排序后，总体保留得分最高的3条加入提示词。对于句向量的获取（$Emb$），我们采用了为检索增强大语言模型而设计的bge-base-zh-v1.5(Chen et al., 2023)。


(c) 2024 中国中文信息学会计算语言学专业委员会

在**模型选择**方面，我们分别选择GPT-3.5与ChatGLM2-6B(Du et al., 2022; Zeng et al., 2022)。其中，GPT-3.5是基于Transformer(Vaswani et al., 2023)的黑盒模型，在多种语言与任务下都取得了良好效果，是最为广泛使用的大型多语言模型之一，本文通过API进行调用。ChatGLM2模型是基于GLM架构的开源双语模型，在预训练阶段就加入了大量的中文数据集，使得其在许多中文领域任务中能够取得显著的效果，我们通过开源模型进行调用。

在**模型设置**方面，为确保实验的公平性和可复现性，所有模型均设定为相同的参数。其中，我们设置温度为0，频率惩罚和top-k为0，top-p为1，以尽量接近贪婪搜索策略。对于无法精确匹配上述参数的ChatGLM2，我们尽可能地调整其设置以接近这些参数。

在**提示词设置**方面，考虑到生成式语言模型对提示词的敏感性，在每个任务下，我们都首先为模型构建3种常用的提示词形式，并指导GPT-3.5额外生成7条表达方式各异、长度不同的新提示词。随后我们通过小批量数据验证的方式从中选取对于当前模型和任务最适宜的，作为后续的评估提示词，以确保最大程度激发模型的能力。具体的提示词设置情况见附录A.2。

特别需要说明的是，对于语素义与词义消歧任务，为了避免检索阶段对预测结果产生标签泄露，我们对所有检索结果都进行了过滤。对于直接提供类似“{目标字}的释义是{标签释义}”或“{目标字}在{目标上下文}中的释义是{标签释义}”的检索结果，我们均将其删除。

## 5.4 语言准确理解任务评估结果与分析

### 5.4.1 主实验结果

我们的方法在三个评估任务上的主要实验结果如表7所示。鉴于本文的核心目标是验证字词资源在大语言模型检索增强生成中的应用效果，我们选择了模型在未注入相关知识时的性能作为参照基准。为了达到更好的覆盖度，本文采用了各任务数据的全集；而在传统方法中，由于需要做模型训练，采用的评估数据集往往是其子集。由于数据覆盖的差异，本文方法与传统方法无法做公平对比。作为参考，传统方法的最佳结果见附录D。其中，对于隐喻识别任务填空形式下的随机基线，其计算结果基于pkuseg分词器获取的上下文中所有可能词汇的数量。

| 模型 | 语素义消歧 | | 词义消歧 | | | 隐喻识别 | | | Avg. |
|---|---|---|---|---|---|---|---|---|---|
| | 词内 | 句内 | 单字词 | 二字词 | 多字词 | 选择 | 判断 | 填空 | |
| **Random** | 15.35 | 12.84 | 18.60 | 44.24 | 45.07 | 25.00 | 50.00 | 21.86 | 29.12 |
| **GPT-3.5** | 58.26 | 46.62 | 50.28 | 62.87 | 67.27 | 36.46 | 56.23 | 53.46 | 53.93 |
| **+本文方法** | **67.93** | **48.89** | **61.57** | **66.59** | **68.40** | **41.29** | **56.43** | **61.39** | **59.06** |
| **ChatGLM2** | 33.82 | 32.48 | 43.67 | 55.80 | 59.92 | 24.66 | 52.84 | 47.65 | 43.86 |
| **+本文方法** | **48.15** | **33.85** | **47.32** | **59.05** | **62.30** | **26.65** | **55.01** | **49.63** | **47.75** |

Table 7: 主实验结果，其中*Random*代表随机基线

表7中的实验成果表明，本文方法在所有实验模型与任务中均带来了显著提升，平均达4.78%。在任务层面，本文方法分别带来了6.91%、4.24%和3.19%的准确率提升，凸显了这一方法在不同任务中的普遍适用性和和有效性。在模型层面，本文方法在GPT-3.5与ChatGLM2上分别提升准确率5.13%、3.89%，这表明无论是面向英文或中文的模型，本文方法都能为其带来有效的知识。其中，GPT-3.5虽然整体效果优于ChatGLM2，但在本文方法下的提升更为显著，展示出这一方法面向更强大模型的潜力。

为了深入探究字词资源的影响，我们对三个任务分别做案例分析，如附录E所示。分析结果表明，在直接注入与目标语素或词相关的知识之外，上下文中其他字词的知识同样有助于优化模型的表现，展示出字词资源广泛的应用价值。这也解释了我们在设置$C_{keyword}$与$C_{sim}$为上下文而非目标语素或词的情况下，依旧能带来显著的提升的原因。

### 5.4.2 消融实验结果

为了验证本文方法中各个模块的有效性，我们对检索、排序及增广例句等关键组成部分进行了消融实验。实验结果如表8所示，其中“排序$_2$”与“排序$_1$”分别代表第二、第一阶段排序。

随排序$_2$、排序$_1$与检索模块的逐步移除，模型在各任务上的准确率均呈递减趋势，证明了它们在获取字词信息上的重要性。其中，移除排序$_2$后，模型准确率分别


(c) 2024 中国中文信息学会计算语言学专业委员会

| 模型 | 方法 | 语素义消歧 | | 词义消歧 | | | 隐喻识别 | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 词内 | 句内 | 单字词 | 二字词 | 多字词 | 选择 | 判断 | 填空 |
| GPT-3.5 | **本文方法** | 67.93 | **48.89** | **61.57** | 66.59 | **68.40** | **41.29** | **56.43** | **61.39** |
| | **−排序$_2$** | 67.63 | 47.02 | 61.56 | 66.36 | 67.55 | 41.17 | 56.40 | 58.15 |
| | **−排序$_{1\&2}$** | 66.82 | 46.99 | 56.88 | 62.11 | 63.82 | 40.98 | 56.31 | 58.05 |
| | **−检索&排序$_{1\&2}$** | 53.70 | 43.74 | 54.08 | 61.99 | 63.75 | 33.14 | 56.29 | 58.00 |
| | **−增广例句** | **68.02** | 47.32 | 61.00 | **67.45** | 67.89 | 41.26 | 56.43 | 58.78 |
| Chat-GLM2 | **本文方法** | **48.15** | **33.85** | **47.32** | 59.05 | **62.30** | **26.65** | **55.01** | **49.63** |
| | **−排序$_2$** | 44.05 | 33.63 | 47.31 | 59.04 | 62.13 | 26.65 | 55.00 | 49.58 |
| | **−排序$_{1\&2}$** | 42.93 | 31.90 | 42.72 | 54.03 | 57.77 | 26.08 | 52.62 | 49.43 |
| | **−检索&排序$_{1\&2}$** | 30.82 | 28.41 | 39.85 | 53.50 | 57.72 | 24.81 | 45.82 | 49.25 |
| | **−增广例句** | 44.70 | 33.14 | 46.77 | **59.56** | 60.83 | 25.72 | 47.38 | 49.38 |

Table 8: 消融实验结果

平均下降0.83%、0.57%；移除排序$_1$下降1.74%、2.59%；移除检索模块则进一步分别带来3.41%、3.41%的下降。

对于字词资源，移除增广例句后两个模型的平均准确率分别下降0.46%、1.81%，印证了其价值。然而，这一效果并非在所有任务中都同样显著。例如，在词内语素义消歧任务中，GPT-3.5在移除增广例句后准确率反而小幅提升了0.09%。

### 5.4.3 例句信息对实验结果的影响

我们注意到，部分场景下采用无增广例句的实验效果略优于增广后的。为了探究其背后的原因，我们以ChatGLM2模型为例，分别分析字词资源内例句长度与覆盖率的影响。

表9展示了例句长度对实验结果的影响，“平均长度”指字词资源中所有例句的平均长度，“覆盖率”指其中包含例句的义项占全部义项的比例，与表4中的“覆盖率”一致。构建不同长度字词资源的方法如下：原始资源中的同一义项下可能包含多条例句。对于所有义项，我们仅选取其中最短的一条，得到的字词资源中平均例句长度为22.23；选取其中长度第二短的例句，得到的字词资源中平均例句长度为31.40，以此类推。结果显示，随例句平均长度的增加，模型在三个任务上的指标均呈下降趋势，这可能是因为较长的例句会分散模型对关键信息的注意力。然而，在单字词义消歧和隐喻识别的填空任务中，较长的例句反而能取得较好的效果，展示出不同任务对其长度的需求存在差异。这一结果也解释了消融实验中，增广例句导致某些任务的效果降低：如表4所示，增广后的字词资源中，例句覆盖率得到了提高，但其平均长度也显著增加了。

| 平均长度 | 覆盖率(%) | 语素义消歧 | | 词义消歧 | | | 隐喻识别 | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 词内 | 句内 | 单字词 | 二字词 | 多字词 | 选择 | 判断 | 填空 |
| 22.23 | 88.00 | **48.97** | **34.11** | 45.94 | **58.84** | **60.77** | **27.16** | **52.92** | 48.51 |
| 31.40 | 88.00 | 48.31 | 34.00 | **46.49** | 58.59 | 60.32 | 26.44 | 52.72 | 49.25 |
| 39.78 | 88.00 | 48.13 | 33.71 | 45.37 | 56.77 | 59.58 | 25.02 | 52.47 | **49.35** |
| 42.21 | 88.00 | 48.25 | 33.68 | 44.80 | 56.58 | 59.52 | 25.32 | 51.69 | 48.91 |
| 43.01 | 88.00 | 48.03 | 32.23 | 44.02 | 56.32 | 59.15 | 24.99 | 51.40 | 48.88 |

Table 9: 例句长度对实验结果的影响

表10展示了例句覆盖率对实验结果的影响。构建不同覆盖率的字词资源的方法如下：第一步，对于所有义项，我们仅保留其中最短的一条例句；第二步，我们随机删除部分义项的例句，构造出平均长度接近但覆盖率不同的字词资源。结果显示，随着例句覆盖率的下降，模型在三个任务上的指标呈下降趋势。这可能是由于过低的例句覆盖率会导致有效信息的缺失，也从侧面证明了例句增广知识的价值。

在例句信息之外，我们还开展了对知识类型与知识数量的实验分析（附录F）：1. 在**知识类型**上，增加词性信息后，两个模型在所有任务上的指标均有下降。这可能是因为释义文本已

| 平均长度 | 覆盖率(%) | 语素义消歧 | | 词义消歧 | | | 隐喻识别 | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 词内 | 句内 | 单字词 | 二字词 | 多字词 | 选择 | 判断 | 填空 |
| 22.23 | 88.00 | **48.97** | **34.11** | **45.94** | **58.84** | **60.77** | **27.16** | **52.92** | **48.51** |
| 22.17 | 70.40 | 48.68 | 33.34 | 45.10 | 58.10 | 60.03 | 26.71 | 52.45 | 48.43 |
| 22.21 | 52.80 | 48.51 | 33.08 | 44.22 | 57.38 | 60.37 | 26.56 | 51.90 | 48.48 |
| 22.14 | 35.20 | 48.47 | 32.52 | 43.73 | 56.77 | 60.03 | 26.08 | 51.54 | 48.20 |
| 22.13 | 17.60 | 48.28 | 32.03 | 43.68 | 56.52 | 60.15 | 26.02 | 51.52 | 48.18 |

Table 10: 例句覆盖率对实验结果的影响

有效地传达了词性相关信息，显式词性标签对大语言模型而言也较难理解。这为未来更多语言学知识的注入提供了参考；2. 在**知识数量**上，模型在$N_{stage1} = 1$时效果更好，侧面反映出一阶段排序的有效性；而$N_{stage2}$的增加会显著提高预测准确率，展示出字词资源对模型的有效性，以及本文方法的潜力。

## 6 结语

本文从语言学角度出发，探索汉语字词资源对大语言模型的语言准确理解能力的效度。基于之前的工作，我们利用网络语料库和生成式模型构建了包含498,126条例句的汉语字词资源，显著提升了其中例句的覆盖率与丰富度。在此基础上，探索这一资源与大语言模型的结合，提出了一套能灵活适应不同任务与模型的检索增强生成方法。评估实验表明，该方法在所有实验模型与任务中均带来了显著的准确率提升，平均达4.78%。其中，在语素义消歧、词义消歧与隐喻识别任务中分别提升了6.91%、4.24%与3.19%，在GPT-3.5与ChatGLM2模型上分别提升了5.22%和4.33%，这显示出字词资源在模型对语言准确理解能力上的价值。通过消融实验，我们进一步验证了这一资源与方法的有效性，并在分析实验中深入探究例句信息、知识类型和知识数量对实验结果的影响，为语言学资源与大语言模型的结合提供新的路径与方法。

在后续工作中，我们将继续开展字词资源的分析与整理，进一步提升例句覆盖率、有效控制其长度。此外，还将探索该资源与方法在更多下游任务与模型中的应用，以便更好地服务于人文与计算领域中的多种任务。

## 参考文献

Eneko Agirre, Oier López de Lacalle, and Aitor Soroa. 2014. Random walks for knowledge-based word sense disambiguation. *Computational Linguistics*, 40(1):57–84.

Kabir Ahuja, Rishav Hada, Millicent Ochieng, Prachi Jain, Harshita Diddee, Samuel Maina, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, et al. 2023. Mega: Multilingual evaluation of generative ai. *arXiv preprint arXiv:2303.12528*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Angels Balaguer, Vinamra Benara, Renato Luiz de Freitas Cunha, Roberto de M. Estevão Filho, Todd Hendry, Daniel Holstein, Jennifer Marsman, Nick Mecklenburg, Sara Malvar, Leonardo O. Nunes, Rafael Padilha, Morris Sharp, Bruno Silva, Swati Sharma, Vijay Aski, and Ranveer Chandra. 2024. Rag vs fine-tuning: Pipelines, tradeoffs, and a case study on agriculture.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2023. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation.


(c) 2024 中国中文信息学会计算语言学专业委员会

Xiaolan Chen, Ziwei Zhao, Weiyi Zhang, Pusheng Xu, Le Gao, Mingpu Xu, Yue Wu, Yinwen Li, Danli Shi, and Mingguang He. 2024. Eyegpt: Ophthalmic assistant with large language models.

Minjin Choi, Sunkyung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. 2021. MelBERT: Metaphor detection via contextualized late interaction using metaphorical identification theories. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1763–1773, Online, June. Association for Computational Linguistics.

Iulia Comșa, Julian Eisenschlos, and Srini Narayanan. 2022. MiQA: A benchmark for inference on metaphorical questions. In Yulan He, Heng Ji, Sujian Li, Yang Liu, and Chua-Hui Chang, editors, *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 373–381, Online only, November. Association for Computational Linguistics.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Michel Galley and Kathleen McKeown. 2003. Improving word sense disambiguation in lexical chaining. In *Proceedings of 18th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1486–1488.

Yuan Gao, Ruili Wang, and Feng Hou. 2023. Unleashing the power of chatgpt for translation: An empirical study. *arXiv preprint arXiv:2304.02182*.

Tiezheng Guo, Qingwen Yang, Chen Wang, Yanyi Liu, Pan Li, Jiawei Tang, Dapeng Li, and Yingyou Wen. 2024. Knowledgenavigator: Leveraging large language models for enhanced reasoning over knowledge graph.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. GlossBERT: BERT for word sense disambiguation with gloss knowledge. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3509–3514, Hong Kong, China, November. Association for Computational Linguistics.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-Eval: A multi-level multi-discipline Chinese evaluation suite for foundation models.

Haoqiang Kang and Xiao-Yang Liu. 2023. Deficiency of large language models in finance: An empirical examination of hallucination.

Minki Kang, Jin Myung Kwak, Jinheon Baek, and Sung Ju Hwang. 2023. Knowledge graph-augmented language models for knowledge-grounded dialogue generation. *arXiv preprint arXiv:2305.18846*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Yucheng Li, Shun Wang, Chenghua Lin, Frank Guerin, and Loic Barrault. 2023. FrameBERT: Conceptual metaphor detection with frame embedding learning. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1558–1563, Dubrovnik, Croatia, May. Association for Computational Linguistics.

CCL 2024

Xiao Liu, Xuanyu Lei, Shengyuan Wang, Yue Huang, Zhuoer Feng, Bosi Wen, Jiale Cheng, Pei Ke, Yifan Xu, Weng Lam Tam, et al. 2023. Alignbench: Benchmarking chinese alignment of large language models. *arXiv preprint arXiv:2311.18743*.

Hongyuan Lu, Haoyang Huang, Dongdong Zhang, Haoran Yang, Wai Lam, and Furu Wei. 2023. Chain-of-dictionary prompting elicits translation in large language models. *arXiv preprint arXiv:2305.06575*.

Fuli Luo, Tianyu Liu, Zexue He, Qiaolin Xia, Zhifang Sui, and Baobao Chang. 2018. Leveraging gloss knowledge in neural word sense disambiguation by hierarchical co-attention. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1402–1411.

Ruixuan Luo, Jingjing Xu, Yi Zhang, Zhiyuan Zhang, Xuancheng Ren, and Xu Sun. 2019. Pkuseg: A toolkit for multi-domain chinese word segmentation. *CoRR*, abs/1906.11455.

Can Luo. 2013. The classification of morphemes, words and its grammatical and semantic categories in the 3000 characters. Master's thesis, Jinan University, Guangzhou, China.

Gonzalo Martínez, Javier Conde, Pedro Reviriego, Elena Merino-Gómez, José Alberto Hernández, and Fabrizio Lombardi. 2023. How many words does chatgpt know? the answer is chatwords. *arXiv preprint arXiv:2309.16777*.

Sheshera Mysore, Zhuoran Lu, Mengting Wan, Longqi Yang, Steve Menezes, Tina Baghaee, Emmanuel Barajas Gonzalez, Jennifer Neville, and Tara Safavi. 2023. Pearl: Personalizing large language model writing assistants with generation-calibrated retrievers. *arXiv preprint arXiv:2311.09180*.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*, 41(2):1–69.

Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. 2024. Fine-tuning or retrieval? comparing knowledge injection in llms.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *arXiv preprint arXiv:2302.00083*.

I.A. Richards. 1936. *The Philosophy of Rhetoric*. Bryn Mawr College. Mary Flexner lectures. Oxford University Press.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for high- (but not low-) resource languages. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore, December. Association for Computational Linguistics.

Jaromir Savelka, Kevin D. Ashley, Morgan A. Gray, Hannes Westermann, and Huihui Xu. 2023. Explaining legal concepts with augmented large language models (gpt-4).

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020. Sensembert: Context-enhanced sense embeddings for multilingual word sense disambiguation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8758–8765, Apr.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy, July. Association for Computational Linguistics.

CCL 2024

Chuandong Su, Fumiyo Fukumoto, Xiaoxi Huang, Jiyi Li, Rongbo Wang, and Zhiqun Chen. 2020. DeepMet: A reading comprehension paradigm for token-level metaphor detection. In Beata Beigman Klebanov, Ekaterina Shutova, Patricia Lichtenstein, Smaranda Muresan, Chee Wee, Anna Feldman, and Debanjan Ghosh, editors, *Proceedings of the Second Workshop on Figurative Language Processing*, pages 30–39, Online, July. Association for Computational Linguistics.

Weisong Sun, Chunrong Fang, Yudu You, Yuchen Chen, Yi Liu, Chong Wang, Jian Zhang, Quanjun Zhang, Hanwei Qian, Wei Zhao, et al. 2023. A prompt learning framework for source code summarization. *arXiv preprint arXiv:2312.16066*.

Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. 2023. A benchmark for learning to translate a new language from one grammar book. *arXiv preprint arXiv:2309.16575*.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Xiaoyu Tong, Rochelle Choenni, Martha Lewis, and Ekaterina Shutova. 2024. Metaphor understanding challenge dataset for llms.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.

Lennart Wachowiak and Dagmar Gromann. 2023. Does GPT-3 grasp metaphors? identifying metaphor mappings with generative language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1018–1032, Toronto, Canada, July. Association for Computational Linguistics.

Yue Wang, Hua Zheng, Yaqi Yin, Hansi Wang, Qiliang Liang, and Yang Liu. 2024. Morpheme sense disambiguation: A new task aiming for understanding the language at character level. Proceedings of the 29th International Conference on Computational Linguistics.

Liner Yang, Cunliang Kong, Yun Chen, Yang Liu, Qinan Fan, and Erhong Yang. 2020. Incorporating sememes into chinese definition modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1669–1677.

Yaqi Yin, Yue Wang, and Yang Liu. 2024. Chinese morpheme-informed evaluation of large language models. Proceedings of the 29th International Conference on Computational Linguistics.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Kexun Zhang, Yee Man Choi, Zhenqiao Song, Taiqi He, William Yang Wang, and Lei Li. 2024a. Hire a linguist!: Learning endangered languages with in-context linguistic descriptions. *arXiv preprint arXiv:2402.18025*.

Qinggang Zhang, Junnan Dong, Hao Chen, Daochen Zha, Zailiang Yu, and Xiao Huang. 2024b. Knowgpt: Knowledge injection for large language models.

Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2024c. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *Advances in Neural Information Processing Systems*, 36.

Kai Zhang. 1997. Statistical analysis of Chinese morpheme-based characters. *Language Teaching and Linguistic Studies*, (01):43–52.


(c) 2024 中国中文信息学会计算语言学专业委员会

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Hua Zheng, Damai Dai, Lei Li, Tianyu Liu, Zhifang Sui, Baobao Chang, and Yang Liu. 2021a. Decompose, fuse and generate: A formation-informed method for Chinese definition generation. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5524–5531, Online, June. Association for Computational Linguistics.

Hua Zheng, Lei Li, Damai Dai, Deli Chen, Tianyu Liu, Xu Sun, and Yang Liu. 2021b. Leveraging word-formation knowledge for chinese word sense disambiguation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 918–923.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. AGIEval: A human-centric benchmark for evaluating foundation models.

Houquan Zhou, Yang Hou, Zhenghua Li, Xuebin Wang, Zhefeng Wang, Xinyu Duan, and Min Zhang. 2023. How well do large language models understand syntax? an evaluation by asking natural language questions.

亢世勇, 李毅, 孙道功, and 张楠. 2004. 汉语系统语料库的建设与词典编纂. *2004*年辞书与数字化研讨会论文集, pages 145–151.

刘扬, 林子, and 康司辰. 2018. 汉语的语素概念提取与语义构词分析. 中文信息学报, 32(2):10.

吉志薇and 冯敏萱. 2015. 面向普通未登录词理解的二字词语义构词研究. 中文信息学报, 29(05):63–68+83.

尹斌庸. 1984. 汉语语素的定量研究. 中国语文, 1(05):338–347.

徐枢. 1990. 语素. 人民教育出版社.

朱德熙. 1982. 语法讲义. 商务印书馆.

梅家驹, 竺一鸣, 高蕴琦, and 殷鸿翔. 1983. 同义词词林. 上海辞书出版社:上海.

符淮青. 1981. 词义和构成词的语素义的关系. 辞书研究, 1:98–110.

苑春法and 黄昌宁. 1998. 基于语素数据库的汉语语素及构词研究. 世界汉语教学, 02:8–13.

郑廻. 2022. 汉语语素构词的嵌入表示与应用评估. Master's thesis, 北京大学.

陈龙, 饶琪, and 刘扬. 2019. 汉语词的非字面义的表示与应用. 中国科学：信息科学, 49(08):1005–1018.

# A 提示词选取结果

## A.1 面向模型增强的汉语字词知识工程提示词

面向模型增强的汉语字词知识工程提示词如表11所示，其中{$w_T$}为目标字或词，{$s_T$}为目标释义。

## A.2 基于字词资源的检索增强评估提示词

三个评估任务下选用的提示词如图4所示。其中词内、句内语素义消歧所用提示词一致，单字词、二字词与多字词词义消歧所用提示词一致。在语素义消歧任务中，考虑到大语言模型对语言学概念不熟悉，数据集内也均为单字语素，我们并未在提示词中采用“语素”，而是用模型更容易理解的“字”作为替换。


(c) 2024 中国中文信息学会计算语言学专业委员会

| 模块 | 提示词 |
|---|---|
| 生成模块 | 请用“{$w_T$}”（{$s_T$}）造句。 |
| | 请用词“{$w_T$}”（释义：{$s_T$}）造句。 |
| | 已知词“{$w_T$}”的意思是“{$s_T$}”，请用词“{$w_T$}”造句。 |
| 检查模块 | 判断句子“{$c_w^i$}”中词“{$w_T$}”是否为“{$s_T$}”的意思。回答是或否。 |
| | 句子“{$c_w^i$}”中词“{$w_T$}”是否可以被解释为“{$s_T$}”？回答是或否。 |
| | 句子“{$c_w^i$}”中词“{$w_T$}”的意思是否是“{$s_T$}”？回答是或否。 |

Table 11: 面向模型增强的汉语字词知识工程中，生成模块与检查模块提示词

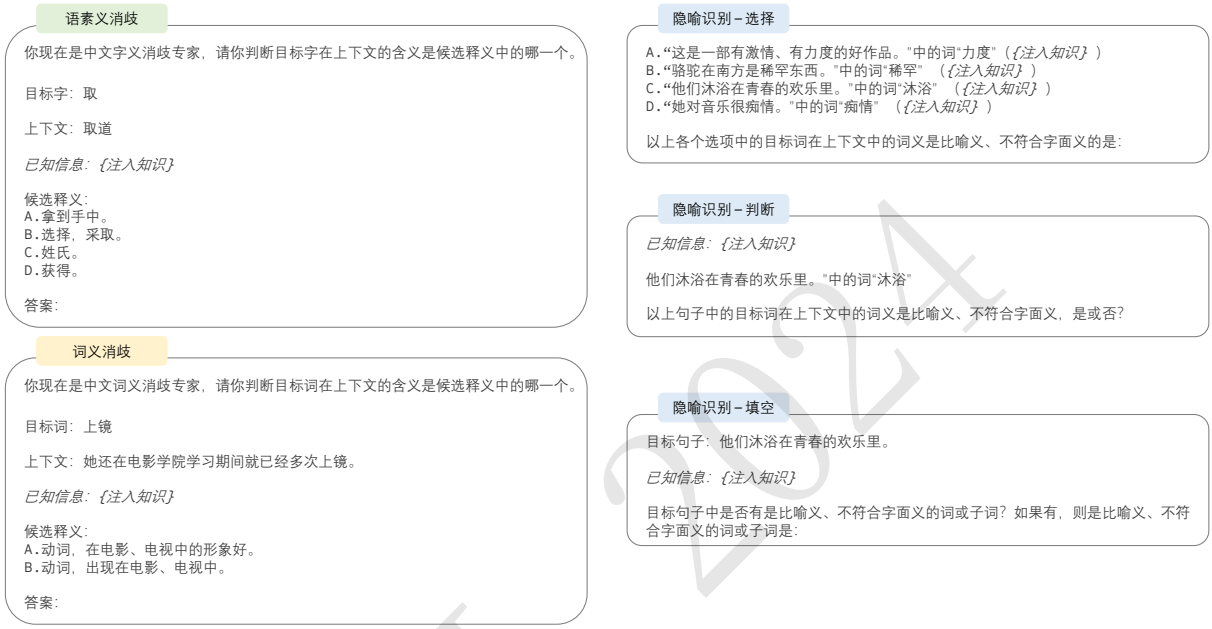



Figure 4: 基于字词资源的检索增强评估提示词

## B 部分语言学专有名词说明

**自由语素**指能够独立构词，无须附于其他语素的语素，例如“花：可供观赏的植物”，可以组词为“花朵”，也可以在句子中独立使用。**黏着语素**与之对应，指不能够独立构词的语素，例如“花：用来迷惑人的；不真实或不真诚的”，只能构成“花招”、“花言巧语”等词语，而不能独立使用。

**语义域**（semantic domain）语义域是一组语义趋近一致的语言表达的集合。隐喻即为从一个域到另一个域的投射。例如，在句子“他们沐浴在青春的欢乐里。”中，词“沐浴”包含隐喻，它的字面义所处的语义域为“洗澡，沐浴，冲凉，浴，擦澡，淋洗，洗浴”，而非字面义所处的语义域为“沉浸，沐浴，正酣”。

## C 主要符号对照表

第4节中所使用的符号说明如表12所示。

## D 传统方法的最佳结果

三项评估任务在传统方法中的最佳结果如表13所示。这些评估在数据覆盖和任务形式等方面与本文存在差异，但大模型在性能指标上显著落后于传统方法。这展示出大模型在此类任务上的不足，也从侧面反映出本文研究的必要性。


(c) 2024 中国中文信息学会计算语言学专业委员会

| 符号 | 说明 |
|---|---|
| $C_{keyword}$ | 用于获取关键词的文本 |
| $C_{sim}$ | 用于计算相似度的文本 |
| $K$ | 关键词集合 |
| $k$ | 关键词 |
| $D$ | 候选知识集合 |
| $d$ | 候选知识 |
| $Emb$ | 获取文本的句嵌入向量 |
| $\alpha$ | 基于评估的权重 |
| $score_{stage1}$ | 一阶段排序分数 |
| $N_{stage1}$ | 一阶段排序得到的知识数量 |
| $score_{stage2}$ | 二阶段排序分数 |
| $N_{stage2}$ | 二阶段排序得到的知识数量 |

Table 12: 主要符号对照表

| 任务 | | 模型 | 测试集准确率 |
|---|---|---|---|
| **语素义消歧** | 词内 | BEM-con+PoS(Wang et al., 2024) | 95.60 |
| | 句内 | BEM-con+PoS(Wang et al., 2024) | 85.47 |
| **词义消歧** | | FormBERT w/FP(Zheng et al., 2021b) | 87.62 |
| **隐喻识别** | | 基于汉语语素的隐喻识别模型(郑颋, 2022) | 81.00 |

Table 13: 传统方法中的最佳结果

## E 案例分析

**语素义消歧**的案例如表14所示，包含词内与句内的语素义消歧案例。其中，释义信息与例句中的释义信息均展示出了显著的效果：通过“世兄”在例句中的释义，模型能够正确消歧“世”；通过“营救”的释义，模型能够正确消歧“营”。这一结果证明了字词信息中的释义与例句中释义两类信息的切实有效性。

| 语素 | 上下文 | 原预测结果 | 注入知识 | 新预测结果 | 标签 |
|---|---|---|---|---|---|
| 世 | 世兄 | E.人世间 | “雷世兄，是我班门弄斧了”中“世兄”的意思是“旧时对辈分相同的世交的称呼” | C.指有世交的关系 | C |
| 营 | 瞿秋白被捕后鲁迅曾想方设法委托人营救 | C.军队编制单位 | “营救”的意思是“设法援救” | B.求，谋求 | B |

Table 14: 语素义消歧任务案例分析

**词义消歧**的案例如表15所示，包含单字词、二字词与多字词的词义消歧案例。在二字词与多字词消歧的案例中，注入知识是上下文内非目标词的释义，但依然有效地辅助了模型对目标词的消歧：对于词“上门”，提供“入赘”的详细释义强化了模型对关键信息的注意力，帮助模型提供了正确的预测结果。这一结果展示出，在目标词之外，上下文中非目标词的知识对模型的语言准确理解能力同样重要且有效。这也解释了我们在主实验中设置$C_{keyword}$与$C_{sim}$为上下文而非目标语素或词，依旧能带来显著提升的原因。

**隐喻识别的选择形式**案例如表16所示。与正确选项相关的知识能够辅助模型直接识别出其隐喻特性，如“膏泽”；而与其他选项相关的知识能够纠正模型错误的语言理解，如“休想”。这一结果展现出，在该任务的选择形式中，对所有选项都提供知识是有助于模型任务表现的。

**隐喻识别的判断形式**案例如表17所示，包含正样本与负样本的案例。其中，提供与目标词相关的或上下文相关的信息均能有效辅助模型预测正确。例如，在“不成”的案例中，有注入知识的情况下，ChatGLM2模型的完整生成结果为：“上下文中，“不成”一词的意思是表示揣测或反问。因此，“不成”一词在上下文中的词义是符合字面义的。”可以看出，正是我们提供的“莫非”的释义信息，辅助了模型对“不成”的理解，因而得到了正确预测结果。这进一步展示出本文

| 词 | 上下文 | 原预测结果 | 注入知识 | 新预测结果 | 标签 |
|---|---|---|---|---|---|
| 交 | 交九的天气 | A.连接；交叉。 | “近日天气渐渐变冷，到了交九的时节。”中“交九”的意思是“进入从冬至开始的‘九’。” | D.到(某一时辰或季节)。 | D |
| 上门 | 入赘俗称「倒插门」、「上门女婿」 | B.到别人家里去；登门。 | “入赘的女婿前几年腰椎损伤”中“入赘”的意思是“男子到女家结婚并成为女家的家庭成员。” | D.指入赘。 | D |
| 包袱底儿 | 这次比赛中他亮出了自己的包袱底儿 | B.指家庭多年不动用的或最贵重的东西。 | “比赛”的意思是“在体育、生产等活动中，比较本领、技术的高低。” | C.比喻最拿手的本领。 | C |

Table 15: 词义消歧任务案例分析

| 提示词 | 原预测结果 | 注入知识 | 新预测结果 | 标签 |
|---|---|---|---|---|
| A.“沙漠一直漫延到遥远的天边。”中的词“漫延”<br>B.“不识字就等于睁眼瞎子。”中的词“等于”<br>C.“膏泽后人。”中的词“膏泽”<br>D.“本位主义”中的词“主义” | D | “膏泽”的意思是“比喻给予恩惠。” | C | C |
| A.“休想逃脱”中的词“休想”<br>B.“下流无耻。”中的词“下流”<br>C.“琅琅上口。”中的词“上口”<br>D.“营造气氛”中的词“营造” | A | “休想”的意思是“别想；不要妄想。” | C | C |

Table 16: 隐喻识别任务选择形式案例分析

方法在多种角度上的有效性，以及对完整上下文应用本文方法的必要性。

| 词 | 上下文 | 原预测结果 | 注入知识 | 新预测结果 | 标签 |
|---|---|---|---|---|---|
| 人梯 | 甘当人梯。 | 否 | “人梯”在目标句子中的词义是“一个人接一个人踩着肩膀向高处攀登叫搭人梯。” | 是 | 是 |
| 不成 | 这么晚他还不来，莫非家里出了什么事不成? | 是 | “莫非”在目标句子中的词义是“表示揣测或反问，常跟“不成”呼应。” | 否 | 否 |

Table 17: 隐喻识别任务判断形式案例分析

**隐喻识别的填空形式**的案例如表18所示。案例展示出，与标签词直接相关的信息能够帮助模型直接意识到其隐喻现象。例如，在“挂钩”的案例中，ChatGLM2模型的完整生成结果为：“目标句子中，“挂钩”一词不符合字面义，因为“挂钩”一词的本义是“用钩子挂住”，与“建立联系”无关。因此，“挂钩”一词是比喻义，不符合字面义。”另一方面，与上下文中其他字词相关的信息也能够帮助模型排除错误的选项，如“远大”的案例。这同样展现出对完整上下文应用本文方法的必要性与有效性。

## F 实验分析

我们以ChatGLM2模型为例，进一步开展了对知识类型与数量的分析实验。

### F.1 注入知识类型对实验结果的影响

主实验中，我们采用了字词资源中的释义与例句信息作为注入知识。在大模型的语言理解中，词性标注（Part-of-Speech, PoS）标签同样被广泛使用(Gao et al., 2023)。因此，我们在此探究更多的注入知识类型对实验结果的影响。

表19中展示了模型在不同知识类型下的实验结果。其中，“例句中的释义”信息简写为“例句”。实验结果表明，词性信息作为注入知识时，在所有任务上带来了准确率的下降。我们考虑这是由于释义文本通常已经有效地传达其词性信息，显式的词性标签对大语言模型相对更难以理解，因此不利于模型的任务表现。


(c) 2024 中国中文信息学会计算语言学专业委员会

| 上下文 | 原预测结果 | 注入知识 | 新预测结果 | 标签 |
|---|---|---|---|---|
| 这两个单位早就挂起钩来了。 | 无 | “挂钩”在目标句子中的词义是“比喻建立某种联系。” | 挂钩 | 挂钩 |
| 前途远大。 | 远大 | “远大”在目标句子中的词义是“长远而广阔，不限于目前。” | 前途 | 前途 |

Table 18: 隐喻识别任务填空形式案例分析

| 知识类型 | 语素义消歧 | | 词义消歧 | | | 隐喻识别 | | |
|---|---|---|---|---|---|---|---|---|
| | 词内 | 句内 | 单字词 | 二字词 | 多字词 | 选择 | 判断 | 填空 |
| + 例句&释义 | **48.15** | **33.85** | **47.32** | **59.05** | **62.30** | **26.65** | **55.01** | **49.63** |
| + 例句&释义&词性 | 47.58 | 32.89 | 45.30 | 53.67 | 60.66 | 25.81 | 52.76 | 48.16 |

Table 19: 注入知识类型对实验结果的影响

## F.2 注入知识数量对实验结果的影响

在主实验中，我们设置$N_{stage1} = 3$，$N_{stage2} = 3$，但不同的注入知识数量可能会对实验结果带来影响。我们以词义消歧任务为例，在这一分析实验中对其进行探究。

| $N_{stage1}$ \ $N_{stage2}$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 46.08 | 46.35 | 47.69 | 48.30 | 48.69 |
| 2 | 46.08 | 46.03 | 47.44 | 48.14 | 48.41 |
| 3 | 46.08 | 46.03 | 47.32 | 48.14 | 48.41 |
| 4 | 46.08 | 46.03 | 47.44 | 48.14 | 48.41 |
| 5 | 46.08 | 46.03 | 47.44 | 48.14 | 47.61 |

(a) 单字词词义消歧

| $N_{stage1}$ \ $N_{stage2}$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 58.64 | 60.06 | 61.12 | 62.11 | 62.80 |
| 2 | 58.64 | 58.65 | 59.07 | 59.51 | 59.68 |
| 3 | 58.64 | 58.65 | 59.07 | 59.51 | 59.68 |
| 4 | 58.64 | 58.65 | 59.07 | 59.51 | 59.68 |
| 5 | 58.64 | 58.65 | 59.07 | 59.51 | 59.68 |

(b) 二字词词义消歧

| $N_{stage1}$ \ $N_{stage2}$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 59.30 | 61.56 | 62.24 | 62.97 | 64.10 |
| 2 | 59.30 | 61.22 | 62.13 | 62.58 | 62.58 |
| 3 | 59.30 | 61.22 | 62.13 | 62.58 | 62.58 |
| 4 | 59.30 | 61.22 | 62.13 | 62.58 | 62.58 |
| 5 | 59.30 | 61.22 | 62.13 | 62.58 | 62.69 |

(c) 多字词词义消歧

Figure 5: 注入知识数量对实验结果的影响

实验结果如图5所示。总体来说，随着$N_{stage1}$的增加模型的指标无显著波动，部分场景下在$N_{stage1} = 1$时取得最佳结果，侧面展现出一阶段排序的有效性；随着$N_{stage2}$的增加模型在词义消歧任务上的指标均显著增加，展示出知识注入在数量上的可扩展性，进一步验证了汉语字词信息对模型的价值。

CCL 2024

# 面向CQL的语料库检索引擎的高效实现

**刘廷超**[1]，**鲁鹿鸣**[1,2]，**杨麟儿**[1,2*]，**王雨**[1]
[1]北京语言大学 信息科学学院
[2]北京语言大学 国家语言资源监测与研究平面媒体中心
liutingchao@hotmail.com

## 摘要

语料库检索工具在语言学研究领域具有举足轻重的地位，对于高效获取信息至关重要。然而，当前国内语料库检索工具在语料库检索语言上缺乏统一标准，尤其支持语料库查询语言（CQL）的中文语料库检索工具相对稀缺。在使用不同分词粒度的语料库工具进行中文语料库检索时，会遇到噪声或数据召回难问题。为应对这些挑战，我们研发了支持多粒度分词的CQL 解析器系统CAMELS：一款支持CQL 语句检索，且兼容多粒度分词，支持非词典词检索的语料库检索引擎。经过多种分词器的测试，该引擎展现出了优异的召回率，并在性能上超越了BlackLab的检索速度，为语言学工作者提供了更加易用、精准的检索工具。

**关键词：** 语料库查询语言 ； 语料库检索 ； 倒排索引

# Efficient Implementation of a CQL-oriented Corpus Retrieval Engine

**Tingchao Liu**[1], **Luming Lu**[1,2], **Liner Yang**[1,2*], **Yu Wang**[1]
[1]School of Information Science, Beijing Language and Culture University
[2]National Language Resources Monitoring and Research Center
Print Media Language Branch, Beijing Language and Culture University
liutingchao@hotmail.com

## Abstract

Corpus retrieval tools play an important role in the field of linguistic research and are crucial for efficient access to information. However, the current domestic corpus retrieval tools lack a unified standard in terms of retrieval language, especially the relative scarcity of Chinese corpus retrieval tools that support Corpus Query Language (CQL). When using corpus tools with different granularity cuts for Chinese corpus retrieval, the noise problem or data recall difficulty problem is encountered. To cope with these challenges, we have developed CAMELS (CQL Analyzer with Multi-lEvel Lexical System). CAMELS supports CQL statement retrieval, is compatible with multi-granularity segmentation, and supports retrieval of words that are not in the vocabulary. After testing with multiple tokenizers, the engine demonstrates excellent recall and outperforms BlackLab's retrieval speed in performance, providing linguists with a more easy-to-use and precise retrieval tool.

**Keywords:** corpus query language , corpus retrieval , inverted index

* 通讯作者
**基金项目**：教育部人文社科青年基金（23YJCZH264）；中央高校基本科研业务费（北京语言大学梧桐创新平台，21PT04）
©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

## 1 引言

随着语言学的深入研究和信息技术的发展，语料库作为语言学研究的基础资源，其重要性日益凸显。语料库检索引擎作为从这一宝贵资源中获取信息的核心工具，对于高效、准确地挖掘语言数据具有关键作用(程学旗et al., 2014)。通过语料库检索引擎，研究者可以迅速地检索到大量真实的语言使用实例，从而对语言的规律性、变异性、历时演变等方面进行深入分析。此外，语料库查询系统还能够辅助语言教学(Boulton, 2017)，为学生提供真实的语言学习材料，帮助其更好地理解和掌握语言知识。在翻译领域，翻译者可以通过对比双语语料库，寻找合适的翻译对等词，提高翻译的准确性和流畅性(Hu et al., 2019; Wieting and Gimpel, 2017)。而在自然语言处理领域，大规模的语料库和高效的查询系统则是训练语言模型、开发智能应用的基础(奚雪峰and 周国栋, 2016)，并可以推进多种NLP下游任务的研究(Chong et al., 2023; Kong et al., 2022)。然而，当前国内语料库检索引擎在技术和应用层面仍存在诸多挑战。

一方面，在各个语料库引擎之间存在检索语言不能通用的问题。现有的一些成功的语料库检索系统，其检索方法往往各不相同，如北京大学的CCL现代汉语语料库（Center for Chinese Linguistics Peking University，CCL）(詹卫东et al., 2019)、北京语言大学的BCC语料库（Beijing Language and Culture University Corpus Center，BCC）(荀恩东et al., 2016)、国家语言资源动态流通语料库（Dynamic Circulation Corpus，DCC）(朱君辉et al., 2022)，他们都拥有丰富的语料资源和多样的检索功能，但各自使用的检索语言差异显著。在检索语句方面，CCL的查询表达式由操作符、基本项、简单项、复杂项、过滤项、子句等构成，定义了13个特殊符号，能够实现相对复杂的检索表达式。BCC具备两套查询表达式，其一为针对序列语料的交互式查询检索式，由查询对象、限制条件和功能操作组成；其二为针对复杂知识挖掘检索需求设计的脚本式查询表达式，同样由三部分组成，查询逻辑通过分解成一系列函数来完成(荀恩东, 2023)。DCC 在检索语言设计上采用了不同的策略，其支持普通检索和模式检索，其中普通检索模式采用了基于语料库查询语言（Corpus Query Language，CQL）的扩展作为检索语言(Valenzuela-Escárcega et al., 2020)。尽管检索表达式的多样性和复杂性体现了语料库检索工具强大的功能，但同时也导致了语料库检索语言的非标准化问题。这意味着用户在使用不同的语料库时，需要学习并掌握各自独特的检索语法和规则。近年来，许多国内外语料库都添加了对CQL 的支持(张永伟and 吴冰欣, 2023)，例如CQPweb(Hardie, 2012)、Sketch Engine(Kilgarriff et al., 2008)、BlackLab (de Does et al., 2017) 等。这是因为CQL语法规则简单，构成的检索式可读性强，例如检索包含“学习”的句子的检索式为：`[word='学习']`，检索包含“学习”且“学习”为动词的句子的检索式为：`[word='学习' & pos='VV']`。

另一方面，语料预处理时是否对语料执行分词会导致不同语料库引擎存在数据噪音或数据召回难问题。例如，在CCL 语料库中，绝大部分查询基于未进行分词和词性标注的纯文本(张永伟and 吴冰欣, 2023)，无法直接用词性检索。CCL 所提供的模式查询功能可以用于限定词语间的搭配模式，然而这种设计会导致检索结果中存在噪音。例如，使用CCL查询语句“爱(X)不(X)”进行语料库查询时，引擎可能会返回如“恩爱得不得了”之类的结果，这是所有基于未分词语料的语料库引擎面临的共性问题，即无法保证用户输入的检索词会被匹配到一个完整的词语，查询结果中存在一定比例的噪音数据。除此之外，在使用分词处理后的语料建立语料库时，由于中文词语的切分粒度不固定且存在一定的歧义性，当用户输入的检索词与分词词表中的词语不一致时，检索系统就无法找到对应的倒排索引信息，从而无法获得期望的结果，数据召回问题由此产生。例如，对于词语“和平共处五项原则”，其在Stanford CoreNLP (Manning et al., 2014)中会被切分成为“和平”“共处”“五”“项”“原则”等词语，而在Jieba[1] 工具中被视为“和平共处”“五项原则”两个词语。THULAC[2]则会把整个“和平共处五项原则”视作一个词语。如果以词语“和平共处”为检索词，在以上述三种分词方式处理的语料构建的语料库中查询，引擎将不会返回任何结果。

针对语料库领域的现存问题，我们开发了一种新的语料库检索引擎——骆驼：**多粒度分词的CQL 解析器系统(CQL Analyzer with Multi-lEvel Lexical System, CAMELS)**[3]。该引擎的主要贡献如下：

(1) 支持使用通用CQL 语法检索。鉴于CQL在语料库领域的广泛应用，本文选择CQL 作为

[1]https://github.com/fxsjy/jieba
[2]http://thulac.thunlp.org/demo
[3]https://github.com/paineliu/camels


(c) 2024 中国中文信息学会计算语言学专业委员会

检索语言，以提升检索语言的通用性。我们基于ANTLR 4 (Parr, 2013) 工具实现了CQL 的编译器，将CQL 语句解析为语法树，进而将语法树转换成包含检索单元、检索范围及约束条件的检索式。这使得语料库检索引擎能够直接利用这些检索式执行语料查询。

(2) 改进了传统倒排索引结构，确保跨分词粒度查询准确性与高效性。在CAMELS引擎中，我们以分词单元为基础，对所有以某个词为起始的语料数据子串进行排序并建立索引。通过让所有索引项直接指向语料数据的方法，避免了额外增加存储空间。同时，CAMELS 引擎基于分组排序合并策略构建整体有序的语料数据索引表。在该索引表中，存储每个词语在语料数据表中的位置。根据词汇表构建的词典TRIE 树保存着索引表中对应词语的起始和结束位置。在用户检索时，引擎基于所输入的检索词的最大前缀在TRIE树中匹配范围，在有序数据索引表中查找，实现了对检索词的匹配文档的快速定位。通过对倒排索引结构的这些优化措施，CAMELS 引擎有效地解决了由分词粒度引起的语料库检索问题。

(3) 基于两阶段检索策略，提升了语料查询效率。在CAMELS引擎中，采用两阶段检索策略。在第一阶段中，引擎根据检索式中的检索范围在倒排索引表上筛选出符合条件的语料；在第二阶段中，引擎则基于给定的约束条件对上一阶段中给出的语料进行验证和进一步筛选。

我们将CAMELS 引擎与其它CQL 语料库检索引擎进行了对比，实验验证了CAMELS 引擎的高效性及对多粒度分词查询的兼容性。我们运用了多种分词器对语料进行分词与词性标注，基于此构建语料库，并使用CAMELS 引擎执行语料检索任务。我们对比了CAMELS 引擎和使用CQL 的开源语料库检索引擎BlackLab。实验结果表明在面对检索词未被包含于语料库分词词表中的情况下，BlackLab无法有效返回候选结果，而CAMELS 引擎能稳定地提供符合预期的候选项。测试结果充分证明了我们的检索引擎在召回率方面有出色表现。

## 2 背景

随着信息技术的发展，语料库查询系统日趋成熟，语料库检索工具也逐渐发展和完善。国内外已有一些语料库广泛地实践了检索工具的可用性。语料库可以按照其分词粒度划分为两类：词级别存储语料库（如DCC 等）和字级别存储的语料库（如CCL等）。然而，由于中文词语边界的模糊性和不确定性，使得用户在中文语料库上进行检索时，会面临有效检索词选择困难的问题。对于词级别存储语料库，用户在不清楚具体词表的情况下可能因为与系统分词不一致而导致查询不到结果；对于字级别存储语料库，用户则难以屏蔽未分词语料带来的检索噪音。这一问题不仅影响了检索的准确性和效率，也制约了现有语料库在中文语料库研究中的进一步应用。因此，如何克服中文词语边界的不确定性，提高中文语料库检索性能，成为当前亟待解决的重要问题。

### 2.1 语料库查询语言

语料库作为语言学研究的重要工具，其查询语言也经历了不断的演变与完善。在现有的语料库查询语言中，CQL 具有重要的地位。CQL 是从语料库查询处理器（Corpus Query Processor，CQP）发展而来的一系列语言，提供了词汇、短语和语法等语言学层面丰富检索功能(吴良平, 2023)，逐渐成为语料库检索领域的主流语言。语料库查询语言(Corpus Query Lingua Franca，CQLF)已经成为国际标准ISO 24631-1 (Bański et al., 2016)，在国际范围内得到广泛认可与应用。CQL语句示例如表1所示。

CQL 语法由检索单元、检索范围、约束条件三个部分组成(Lu et al., 2024)：

(1) **检索单元**。检索单元实现对一个词语的检索，它是构成检索式的基本组成部分。通过具体检索表达式，我们可以产生针对特定属性的查询结果。通过在指代单个词语的中括号中添加“属性名=属性值”对，检索单元可以针对某项属性检索特定词语。例如，`[word='学习']` 就表示查询所有包含“学习”这个词的语料。检索单元内的限定条件允许以逻辑运算符连接以表达对多个属性的综合检索。例如，表达式`[word='学习' & pos='VV']`查询包含“学习”作为动词的语料。检索式中的属性值也允许使用正则表达式，这使得用户能够进行更为灵活的模糊查询。

连续排列的检索单元可以表达检索空间上连续的多个词语，从而满足诸如连续短语等更为复杂的查询需求。未添加属性限制的中括号对`[]`则表示任意的词语。通过在检索单元后添加正则表达式量词可以实现非连续型短语或非连续且位置发生变异的短语(吴良平, 2023)。

(2) **检索范围**。CQL 允许使用within 关键字以限制语料检索的范围。在within 关键字前的部分由检索单元指定，在其后接续XML 标签或能够指代更大范围的检索单元序列来指定检


(c) 2024 中国中文信息学会计算语言学专业委员会

索式的匹配的文档数据范围。XML 标签的内容取决于语料中的定义。在默认情况下，可以使用<s/>表示将检索结果限定在一个句子内。检索范围语法可以和其它查询语句同时使用。

(3) **约束条件**。在约束条件语法中，用户以类似变量声明的形式为检索单元赋予名称，然后通过在检索式末尾添加"::"及后面的约束条件的方式，以此指定检索单元之间的逻辑关系。例如，在CQL语句："`A:[][]B:[] :: A.pos = B.pos`"中，分别为第一个和第三个检索单元定义了名称A和B，并且添加了第一单元和第三单元的词性必须相同的约束条件。

| 功能 | 示例 | 说明 |
|---|---|---|
| 检索单元 | [word='学习'] | 包含"学习"的语料 |
| | [word='编辑' & pos='VV'] | 包含"编辑"作为动词的语料 |
| | [word='热爱'][word='学习'] | "热爱""学习"连续出现的语料 |
| | [word='不但'][]{1,5}[word='而且'] | "不但"和"而且"间隔5字内的语料 |
| 检索范围 | [word='科技']within <s / > | 在同一句中包含"科技"的语料 |
| 约束条件 | A:[] [] B:[][] :: A.word = B.word | 包含形式如"有步骤有计划"语料 |

表 1: CQL语句示例

由于CQL在语法简洁性、直观性、灵活性以及扩展性等方面的优势，使用CQL作为语料库检索的主要语言，不仅能够降低学习成本，提高检索效率，还能够更好地满足用户的语料库检索需求。

## 2.2 倒排索引

倒排索引作为一种高效的信息检索手段，在现代信息处理和搜索引擎中发挥着关键作用(Babenko and Lempitsky, 2014; 刘兴宇, 2004)。它的主要功能是记录词项在文档中的出现情况，建立词项反向映射到文档的索引关系(Schütze et al., 2008)，从而提高搜索效率和匹配度。

在语料库设计中，倒排索引的构建是关键环节。语料库系统会根据所采用的词语、词性等标签，建立这些标签与文档之间的倒排索引信息，从而通过标签信息快速检索到含有这些标签的文档。倒排索引由两个部分组成：词项词典和倒排列表。词项词典由文档集合中出现的所有词项组成。文档ID用来唯一标识一个文档。每个词项对应一个或多个由文档ID以及词频、位置等信息组成的倒排列表，记录了该词项出现过的文档、以及在文档中的出现频率和出现位置等信息。这种索引方式的优点在于其高效性和灵活性，使得系统能迅速响应大量查询请求。具体来说，构建倒排索引包括分词、词性标注及建立索引表等步骤，随后系统会根据这些信息创建倒排索引表，记录标签与文档位置的对应关系。在检索时，用户的查询请求被解析为关键词，系统再利用已建立的倒排索引迅速查找到这些关键词对应的文档ID，从而快速定位并返回相关文档内容。

倒排索引在全文搜索领域应用广泛(Wu et al., 2013; Lin, 2009; Fontoura et al., 2006)，Lucene(Białecki et al., 2012)[1]是全文搜索领域的杰出代表，它是一套用于全文检索的高性能、可伸缩的信息检索工具包。Lucene以其强大的文本分析、索引构建和查询处理能力，为开发者提供了灵活且高效的搜索解决方案。Lucene的核心在于其倒排索引技术，能够快速地定位到包含特定关键词的文档(Gospodnetic et al., 2010)。通过分词、停用词过滤、词干提取等步骤，Lucene能够将原始文本转化为适合搜索的索引结构。Lucene在全文搜索领域有着举足轻重的地位。尽管Lucene 检索功能强大，但并不能支撑语料库检索的词语间距离、词性约束、以及语法约束等复杂检索条件的需求，无法直接用于语料库检索。然而基于Lucene优秀的性能，仍有许多语料库检索工具基于Lucene引擎构建(Elasticsearch, 2018; Shahi and Shahi, 2015)。

## 2.3 BlackLab

BlackLab[2] 是基于Lucene 引擎开发的一款专为语言学家设计，同时适用于历时研究和知识提取等领域的文本检索和分析工具。作为一款开源软件，BlackLab 基于Lucene 实现了强大的检索功能和灵活的使用方式。它能够索引标注文本，支持复杂模式的快速搜索，并可通过配置

[1]https://lucene.apache.org
[2]https://github.com/INL/BlackLab


(c) 2024 中国中文信息学会计算语言学专业委员会

实现个性化索引。此外，BlackLab 还提供了丰富的结果处理功能，如结果集分组与排序、检索单元突出显示等，使得用户能够更高效地分析和利用搜索结果。BlackLab 采用CQL 作为其语料库查询语言。

BlackLab 要求输入语料必须以词级别存储，这导致了在BlackLab 中使用CQL 检索时的分词粒度选择困惑问题。在构建中文语料库时，需要对语料数据进行分词处理。而程序的分词结果往往和人类认知并不一致。这种差异也导致用户在语料库检索时，难以确定一段文字在语料库中的具体切分形式。这使得用户在选择检索词时感到困惑(孙茂松, 1999; 詹卫东et al., 2019)。这种检索词的不确定性给使用者带来了极大的不便，不仅影响了检索的效率和准确性，还可能导致遗漏重要的信息。由于BlackLab 严格以词语边界作为检索条件，因此在使用BlackLab进行中文语料库检索时，用户只有在选择与语料数据切分相一致的检索词的情况下，才可能获得正确的检索结果。中文词汇的丰富性和切分单元的不确定性导致分词工具之间出现显著差异，使得如何选择语料库检索词问题更为突显(张文静et al., 2019)。表2以句子“在和平共处五项原则的基础上”为例，展示了Stanford CoreNLP、THULAC（THU Lexical Analyzer for Chinese）[1]和Jieba 等分词工具的结果，反映出中文词语的边界存在着模糊性。

| 分词工具 | 分词结果 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Stanford | 在 | 和平 | 共处 | 五 | 项 | 原则 | 的 | 基础 | 上 |
| THULAC | 在 | 和平共处五项原则 | | | | | 的 | 基础上 | |
| Jieba | 在 | 和平共处 | | 五项原则 | | | 的 | 基础 | 上 |

表 2: 分词工具结果对比。常用的中文分词工具也会因为训练语料与分词策略的不同，对同一句子产生不同的分词结果。可以看到分词一致性比较低，Stanford CoreNLP 工具包更倾向于切分出较小的单元，清华大学THULAC的分词粒度较大，而Jieba 分词工具的分词粒度适中。

## 3 我们的工作

我们选用CQL作为语料库的检索语言，基于C++、Python开发了高效的语料库检索引擎系统CAMELS。我们借助ANTLR 4 实现对CQL的高效精确解析；对传统的倒排索引策略进行了改进，以解决非词典词汇的匹配难题；还优化了存储结构和检索性能，通过实施两阶段检索策略，显著提升了检索速度。表3对比了CAMELS和传统语料库的异同。语料库的索引生成与检索流程如图1所示。

| 划分方法 | 噪音数据 | 存储空间 | 非词典词查询 | 有效查询词 |
|---|---|---|---|---|
| 基于字 | 高 | 高 | ✔ | 五项原则/五项原/项原则/五项/项原/原则/五/项/原/则/ |
| 基于词 | **无** | **低** | ✗ | 五项原则 |
| 我们的方法 | 低 | **低** | ✔ | 五项原则/五项原/五项/原则/五/原 |

表 3: CAMELS与传统语料库基本单位划分粒度的对比图。以“五项原则”为例，有效查询词是指能够召回“五项原则”所在句的查询词。以字为单位划分语料会产生意外的有效检索词，而以词为单位划分则会导致过于严格的查询词限制。CAMELS较好地平衡了两种方式的优点，并在其它方面有良好表现。

### 3.1 CQL语法解析

我们基于ANTLR 4 实现了CQL 解析器，该解析器接收CQL 作为输入，分析其语义，生成语法树，将语法树转换为CAMELS 引擎支持的结构化检索表达式。ANTLR 4（ANother Tool for Language Recognition 4）[2] 是一款功能强大的开源语法分析工具，允许开发者使用扩展巴斯科范式（Extended Backus–Naur Form，EBNF）来描述语法(Parr et al., 2014)。我们

[1]https://github.com/thunlp/THULAC-Python
[2]https://www.antlr.org


(c) 2024 中国中文信息学会计算语言学专业委员会

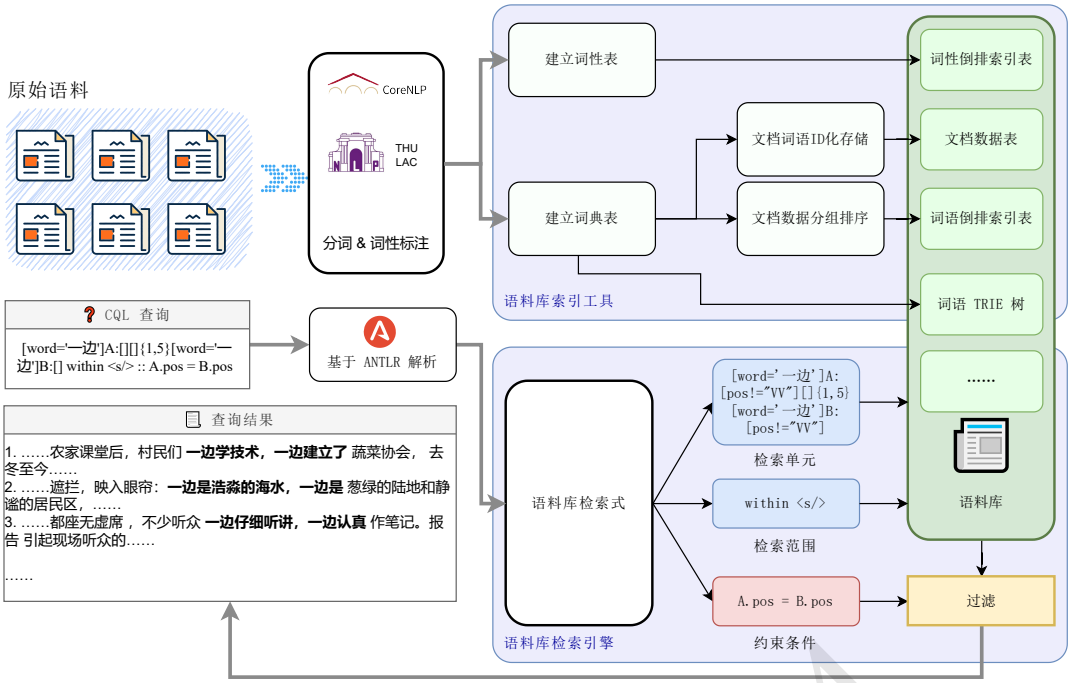



图 1: 语料库索引生成与检索流程图。输入语料被CAMELS 索引工具处理后转换为多个数据表。输入的CQL查询则被转换为CAMELS 内部检索表达式以查询相应语料。图中用CQL 的语句片段表达拆分后的检索表达式的含义，在实际引擎中，查询语句的语义被以另外的形式表述并执行。

在Corpus Query Language Parser开源项目[1] 提供的CQL语言描述文件的基础上，使用ANTLR 4生成了CQL的语法解析器，并基于ANTLR Listener 遍历语法树以实现语义分析和结构化查询表达。分析器构建的一个语法树样例如图2所示。

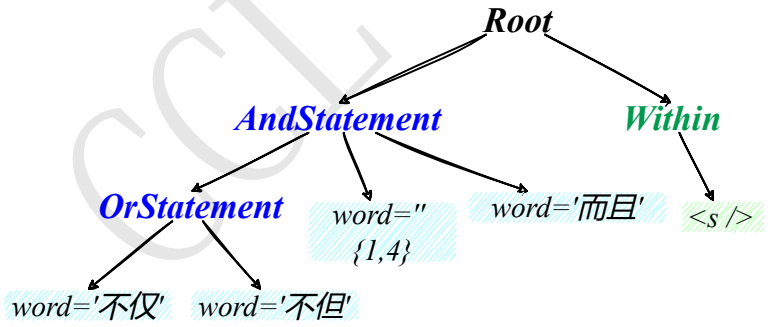



图 2: CQL语法树示意图。以CQL 语句“[word=‘不仅’ |word=‘不但’] []{1,4} [word=‘而且’] within <s/ >”为例，ANTLR生成的语法树结构将该输入语句按层次解析。在这之后，我们基于ANTLR Listener 遍历语法树并完成语义分析。

CQL 经由解析器构建出CAMELS 检索表达式的三个核心组成部分：查询语句、约束条件以及检索范围。查询语句是构成检索表达式的基础元素，它们对应于CQL查询语句中的检索单元。这些查询语句是可以执行的具体查询指令，能够直接对语料库进行数据检索。范围约束由CQL中的检索范围所指定的，其限定了检索单元在语料库中的匹配范围，用户可以指定的范围是句子范围、段落范围或者文档范围。约束条件是由CQL查询语句中的检索条件转换而来的，在检索表达式中起着至关重要的作用。约束条件确保了检索单元之间的关系符合用户的查

[1]https://github.com/exquery/corpusql-parser


(c) 2024 中国中文信息学会计算语言学专业委员会

询意图，它们对检索单元进行逻辑上的连接和限制，从而保证了检索结果满足逻辑关系。由解析器构建出的检索表达式可以被CAMELS 引擎直接运行。

### 3.2 倒排索引表设计

语料库索引工作主要是创建倒排索引表。通常建立语料库索引需要以下步骤：原始文本经过统一编码、段落分割、句子分割等预处理流程后，形成统一格式的语料数据；之后使用语料标注工具或人工标注进行分词和必要信息标注，最后通过语料库索引工具创建倒排索引表以及其它数据表(Kaushik et al., 2004)。由于不同的检索标签的检索需求存在差异，因此需要分别设计倒排索引策略。对于词性检索来说，其检索词是固定且词性体系，因此对于词性采用传统的倒排索引结构设计，直接建立词性到文档之间的索引。

对于词语检索来说，由于中文词语边界的模糊性，需要解决非语料库词表词语的检索匹配问题，因而无法使用传统的倒排索引方式建立索引。词语检索需要具有跨词语边界的检索能力，即对词语倒排索引表的访问需要打破词语边界。例如，对于表2中的例子，用户可能会将“和平共处”（在语料库中是一个完整词语）拆分为“和平”“共处”进行搜索，或者将“五项”（在语料库中是分开的“五”和“项”）组合成一个词进行搜索。此外，在建立倒排索引时，还需要根据语料库提供的范围约束条件，建立文档、段落、句子等多级倒排索引结构。

为实现跨词语边界的检索能力，我们设计了四步流程来构建词语倒排索引表：（1）进行文档词语的ID化转换。我们使用多种分词工具对语料进行分词，保留多种粒度切分，生成词典表，进而将文档内容转换为词语ID序列。（2）对文档数据进行分组排序。文档转换为词语ID序列后，遍历句子中的每个词语，并以此词语为起始点，后续第一个非中文字符为结束点，确定一个子句，然后将这个子句加入到待排序子句列表中。此列表仅记录子句在全部文档中的偏移位置，并不产生新的句子实例。由于待排序子句的数量众多，我们采用句首字符进行分组排序，以提高处理效率。（3）创建词语倒排索引表。当待排序子句完成排序后，会生成一张从语料库中任意词开始的子句排序列表。这使得我们可以利用二分查找法在这张表上进行字符串的匹配范围查询。CAMELS 引擎将子句作为整体字符串与检索词进行比较，而非以词语为单位，从而实现了即便检索词不在词表中也能进行有效检索。最终，我们将子句排序索引以倒排索引表的形式存储，此表实现了从子句到文档的映射。（4）生成词语TRIE树，建立词语到倒排索引表的映射关系。TRIE 树以每个字符为节点进行构建，每个节点下都存储了当前节点在倒排索引表中的匹配范围。在查询关键词的匹配范围时，如果关键词与TRIE树的节点完全匹配，则直接使用匹配节点的检索范围；如果未完全匹配，我们则在TRIE树与检索词的最长匹配节点的范围内进行二分查找，以确定搜索关键词的精确匹配范围。

检索过程的示例如图3所示。当用户搜索“和平共处五项原则”时，检索引擎首先在词典TRIE树上进行匹配，找到最长匹配节点“和平”，并获取其在倒排索引表中的检索范围。由于检索词并未完全匹配，CAMELS引擎需在该范围内进行二分查找，以进一步确定“和平共处”的检索范围，进而将其范围内的候选项与约束条件进行匹配过滤，以获得最终的检索结果。

### 3.3 两阶段检索策略

我们采用两阶段检索策略以保证查询效率与精确性。语料库检索不仅需要实现检索单元的精确匹配，还需验证各检索单元在语料中的位置和距离关系是否符合预设约束。在处理多段检索单元时，先筛选出符合匹配条件的语料，再进行约束条件的验证，可以显著提升检索效率。在CAMELS 检索过程中，我们设置如下两个阶段：

（1）执行检索单元的匹配。在本阶段，引擎执行忽略位置和距离约束的检索表达式，从语料库中检索出所有可能的匹配语料。对于简单检索式，直接查询对应的倒排索引表，返回匹配结果。对于包含逻辑运算的复合检索式，引擎首先执行复合检索式中的简单检索式并取得匹配结果，之后再根据简单检索式之间的逻辑运算关系进行运算，以获得复合检索式的匹配结果。

（2）根据约束条件对第一阶段获取的语料进行筛选。具体而言，引擎依据位置和距离等约束条件，对上一阶段返回的匹配文档进行约束匹配，返回相应的详细信息。引擎从文档的开始位置，逐一核查检索单元的匹配位置是否符合预设的约束条件。由于一个文档内可能存在多个与检索单元相匹配的位置，因此采用动态规划算法，从起始检索单元开始，逐个检查其匹配位置是否满足过滤条件。只有当所有检索单元都符合要求时，我们才将该文档添加到最终的候选列表中。


(c) 2024 中国中文信息学会计算语言学专业委员会

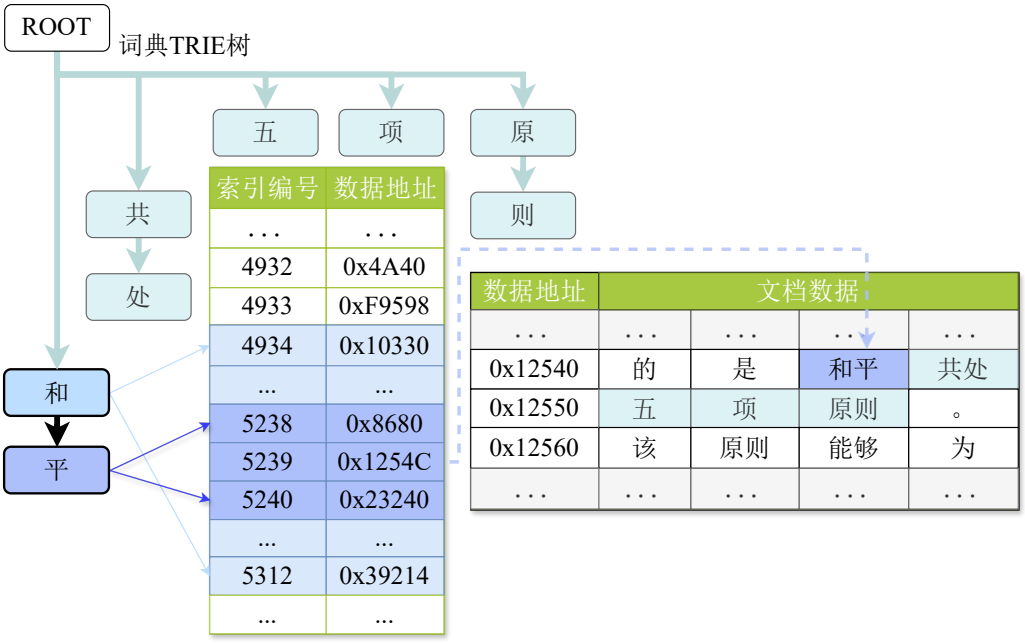



图 3: 倒排索引检索示意图

值得注意的是，合并不同检索单元以及复合检索式内部的简单检索结果时，传统的倒排索引表存储有序文档ID列表，可以通过跳表操作快速合并列表；然而，在我们的方法中，由于词语不是固定的索引单位，且对应的文档ID无序，因此无法采用传统的合并方式。为解决这个问题，我们采用了标记位匹配法，建立一个与文档数量相同的位数组，其中每位表示对应文档是否与检索词匹配。在合并两个检索表达式的结果时，只需对匹配标记进行逻辑运算，即可迅速完成合并操作。

## 4 实验

在本章，我们详细介绍了实验使用的数据与方法。在4.1节，我们介绍了实验使用的数据集。在4.2节，本文对比了BlackLab 与CAMELS 的检索词召回率。在4.3节，本文对比了BlackLab与CAMELS的检索速度。

### 4.1 数据集

实验所采用数据集为《人民日报》1946年至2022年的新闻文本语料。对语料分别使用CoreNLP以及THULAC 分词工具进行处理，再使用BlackLab 与CAMELS 引擎分别对处理过的文件建立索引，生成语料数据库。我们采用人工编撰和机器生成结合的方式编写CQL测试用例。我们基于随机抽取的语料中的句子，自动生成并人工核验了CQL 测试语句，在BlackLab以及CAMELS引擎上进行检索词召回测试和检索速度测试。

### 4.2 检索词召回测试

我们从《人民日报》语料库中随机抽取1000条长度在5至20个汉字的句子，并应用Jieba分词工具进行分词。构造CQL测试用例时，我们首先从句子中随机选取部分词长大于1的词语$W_i$，并以“`[word='`$W_i$`']`”格式生成测试用例。如果该用例无法在相应引擎中返回指定的来源句子，则认为召回失败。我们对于这些构造的CQL测试用例，分别在使用CoreNLP 工具和THULAC 工具建立的语料库上进行检索词的召回测试。召回测试的样例在表4中阐明。实验结果见表5。经测试，CAMELS在所有测试用例上达到了100%的召回率，实验结果表明CAMELS 引擎可以实现跨分词粒度的检索。

分词不一致性是导致召回失败的关键因素。在分析CoreNLP和THULAC分词工具在处理特定测试用例（特别是召回失败案例）中的分词行为后，可以明确识别出这种不一致性：分词工具对语料库的处理结果与检索词之间存在的显著差异。以“至关重要”为例，CoreNLP将其分

| 检索目标 | CQL | BL-S | BL-T | CAMELS |
|---|---|---|---|---|
| 完善农村**宅基地**管理制度。 | [word='宅基'][word='地'] | ✔ | ✗ | ✔ |
| | [word='宅基地'] | ✗ | ✔ | ✔ |
| 这说明扩大内需还要**加把劲**。 | [word='加把劲'] | ✔ | ✗ | ✔ |
| | [word='加'][word='把'][word='劲'] | ✗ | ✔ | ✔ |
| 两个因素**至关重要**。 | [word='至关重要'] | ✗ | ✔ | ✔ |
| | [word='至关'][word='重要'] | ✔ | ✗ | ✔ |
| 我只能**干着急**。 | [word='干着急'] | ✗ | ✔ | ✔ |
| | [word='干'][word='着急'] | ✔ | ✗ | ✔ |
| 这些都是**未解之谜**。 | [word='未解之谜'] | ✗ | ✗ | ✔ |

表 4: 检索词召回实验的示例表。BL-S列代表使用Stanford CoreNLP分词建立的BlackLab语料库，BL-T代表使用清华THULAC分词建立的BlackLab语料库，CAMELS表示本文的语料库。✔表示该CQL在其对应语料库中可以成功召回，✗则表示召回失败。

割为“至关”和“重要”两个独立词语，而THULAC则将其视为一个完整的词语“至关重要”。在处理“未解之谜”时，CoreNLP将其分割为“未解”“之”“谜”三个独立词语，而THULAC则将其分割为“未”“解”“之”“谜”四个独立词语。这种分词差异导致了在现有检索系统中使用用户期望的的检索词检索时，可能会出现无法准确召回目标信息的问题。CAMELS凭借其多粒度分词支持和跨词语检索功能，有效解决了分词不一致性导致的数据召回问题，实现了100%的召回率，显著提升了检索的准确性和效率。

### 4.3 检索速度测试

检索速度测试数据包含50条人工编写的测试用例，和450条自动生成的从人民日报中随机抽取的多字词测试用例，合计500条。我们参考Lu 等(2024) 提出的CQL 生成方法编写检索速度查询测试用例。我们基于汉语增强依存(Yu et al., 2022)抽取搭配并生成检索单元，并随机生成多字词或多字词与其词性的组合的查询语句。针对每一条测试用例，分别应用BlackLab（语料基于Stanford分词）与CAMELS进行12次查询，为确保数据的准确性，我们排除了第一次查询的结果（因为BlackLab在进行第一次查询时会加载数据，导致查询速度较慢），也去除了运行时间为0的异常数据。对剩余的记录取平均值得到最终的结果。实验表明，CAMELS 引擎平均响应时间为13.57毫秒，BlackLab平均响应时间为14.73毫秒，CAMELS 引擎的检索速度相较BlackLab 有7.88% 的提升。实验结果如表5所示。

| 性能 / 系统 | RT(ms) | BR(%) | Recall |
|---|---|---|---|
| CAMELS | **13.57** | **58.2** | **100** |
| BlackLab | 14.73 | 41.8 | 89.6(S) / 93.4(T) |

表 5: 检索速度对比表。RT指平均响应时间，BR指更优比例，表示在所有测试中取得更优结果的数量占总测试数的比例。Recall指在检索词召回实验中的召回比例。表格分别给出了用在Stanford CoreNLP工具分词语料上建立的语料库（S）和在THULAC工具分词语料上建立的语料库（T）检索的结果。CAMELS在召回率上做到了100%。

## 5 结论

本文深入探讨了中文语料库检索工具在检索语言通用性和非词典词匹配召回方面的问题，并针对这些问题设计研发了一款新型的语料库检索引擎。在解决检索语言通用性问题时，我们采用了国际上广泛使用的CQL语言作为语料库的检索语言，并通过ANTLR工具将CQL解析为语法树，实现了高效的检索表达式解析。在解决非词典词匹配召回问题时，我们改进了传统的倒排索引结构，构建了整体有序的语料数据索引表，优化了词语检索匹配策略，实现了非词表词语的有效匹配。同时，我们采用了二阶段检索策略和候选标记位过滤方案，显著提高了检索


(c) 2024 中国中文信息学会计算语言学专业委员会

速度。实验结果显示，与业界知名的BlackLab引擎相比，我们的检索引擎在非词典词检索时，仍能返回预期结果，且在检索速度方面具有明显优势。综上所述，本研究通过自研查询引擎实现了非词典词的检索，并展示了CQL的简洁性和直观性，为语料库检索技术的研究和应用提供了有益参考和高效工具。

## 参考文献


Artem Babenko and Victor Lempitsky. 2014. The inverted multi-index. *IEEE transactions on pattern analysis and machine intelligence*, 37(6):1247–1260.

Piotr Bański, Elena Frick, and Andreas Witt. 2016. Corpus Query Lingua Franca (CQLF). In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2804–2809, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Andrzej Białecki, Robert Muir, Grant Ingersoll, and Lucid Imagination. 2012. Apache lucene 4. In *SIGIR 2012 workshop on open source information retrieval*, page 17. sn.

Alex Boulton. 2017. Corpora in language teaching and learning. *Language Teaching*, 50(4):483–506.

Ruining Chong, Luming Lu, Liner Yang, Jinran Nie, Shuhan Zhou, Yaoxin Li, and Erhong Yang. 2023. Mcts: A multi-reference chinese text simplification dataset.

Jess de Does, Jan Niestadt, and Katrien Depuydt. 2017. Creating research environments with blacklab. *CLARIN in the Low Countries*, pages 245–257.

BV Elasticsearch. 2018. Elasticsearch. *software], version*, 6(1).

Marcus Fontoura, Ronny Lempel, Runping Qi, and Jason Zien. 2006. Inverted index support for numeric search. *Internet Mathematics*, 3(2):153–185.

Otis Gospodnetic, Erik Hatcher, and Michael McCandless. 2010. *Lucene in action*. Simon and Schuster.

Andrew Hardie. 2012. Cqpweb—combining power, flexibility and usability in a corpus analysis tool. *International journal of corpus linguistics*, 17(3):380–409.

J Edward Hu, Rachel Rudinger, Matt Post, and Benjamin Van Durme. 2019. Parabank: Monolingual bitext generation and sentential paraphrasing via lexically-constrained neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6521–6528.

Raghav Kaushik, Rajasekar Krishnamurthy, Jeffrey F Naughton, and Raghu Ramakrishnan. 2004. On the integration of structure indexes and inverted lists. In *Proceedings of the 2004 ACM SIGMOD international conference on Management of data*, pages 779–790.

Adam Kilgarriff, Pavel Rychly, Pavel Smrz, and David Tugwell. 2008. The sketch engine. *Practical Lexicography: a reader*, pages 297–306.

Cunliang Kong, Yun Chen, Hengyuan Zhang, Liner Yang, and Erhong Yang. 2022. Multitasking framework for unsupversied simple definition generation.

Jimmy Lin. 2009. Is searching full text more effective than searching abstracts? *BMC bioinformatics*, 10:1–15.

Luming Lu, Jiyuan An, Yujie Wang, Liner Yang, Cunliang Kong, Zhenghao Liu, Shuo Wang, Haozhe Lin, Mingwei Fang, Yaping Huang, and Erhong Yang. 2024. From text to cql: Bridging natural language and corpus search engine. *ArXiv*, abs/2402.13740.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

Terence Parr, Sam Harwell, and Kathleen Fisher. 2014. Adaptive ll (*) parsing: the power of dynamic analysis. *ACM SIGPLAN Notices*, 49(10):579–598.

Terence Parr. 2013. The definitive antlr 4 reference. *The Definitive ANTLR 4 Reference*, pages 1–326.



(c) 2024 中国中文信息学会计算语言学专业委员会

Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2008. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge.

Dikshant Shahi and Dikshant Shahi. 2015. Apache solr: an introduction. *Apache Solr: A practical approach to enterprise search*, pages 1–9.

Marco A Valenzuela-Escárcega, Gus Hahn-Powell, and Dane Bell. 2020. Odinson: A fast rule-based information extraction framework. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2183–2191.

John Wieting and Kevin Gimpel. 2017. Paranmt-50m: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. *arXiv preprint arXiv:1711.05732*.

Hao Wu, Guoliang Li, and Lizhu Zhou. 2013. Ginix: Generalized inverted index for keyword search. *Tsinghua Science and Technology*, 18(1):77–87.

Jingsi Yu, Shi Jialu, Liner Yang, Dan Xiao, and Erhong Yang. 2022. Transformation of enhanced dependencies in chinese. In *Proceedings of the 21st Chinese National Conference on Computational Linguistics*, pages 99–109.

刘兴宇. 2004. 基于倒排索引的全文检索技术研究. Ph.D. thesis, 武汉: 华中科技大学.

吴良平. 2023. Cqp语法赋能语言研究及语言学习. 语料库语言学, 10:98–114.

奚雪峰and 周国栋. 2016. 面向自然语言处理的深度学习研究. 自动化学报, 42(10):1445–1465.

孙茂松. 1999. 谈谈汉语分词语料库的一致性问题. 语言文字应用, (2):90–93.

张文静, 张惠蒙, 杨麟儿, and 荀恩东. 2019. 基于lattice-lstm的多粒度中文分词. 中文信息学报, 1:7.

张永伟and 吴冰欣. 2023. 基于网络的第四代语料库分析工具核心功能评介. 当代语言学, 25(04):611–624.

朱君辉, 刘鑫, 杨麟儿, 师佳璐, and 杨尔弘. 2022. 文心语料库检索平台的研制. 第十二届全国语言文字应用学术研讨会.

程学旗, 靳小龙, 王元卓, 郭嘉丰, 张铁赢, and 李国杰. 2014. 大数据系统和分析技术综述. 软件学报, 25(9):1889–1908.

荀恩东, 饶高琦, 肖晓悦, and 臧娇娇. 2016. 大数据背景下bcc语料库的研制. 语料库语言学, 3:93–109+118.

荀恩东. 2023. 自然语言结构计算*BCC*语料库. 自然语言结构计算BCC语料库.

詹卫东, 郭锐, 常宝宝, 谌贻荣, and 陈龙. 2019. 北京大学ccl语料库的研制. 语料库语言学, 6(01):71–86+116.

CCL 2024

# NNP-TDGM: 基于最近邻提示表征的术语DEF生成模型

**沈思嘉　王裴岩*　王胜任　王立帮**
沈阳航空航天大学计算机学院，辽宁沈阳110136
沈阳飞机工业（集团）有限公司信息化中心，辽宁沈阳110003
wangpy@sau.edu.cn　wangsr002@avic.com
{shensijia,wanglibang}@stu.sau.edu.cn

## 摘要

该文研究基于HowNet的知识库描述语言语法体系的术语DEF自动生成问题，提出基于最近邻提示表征的术语DEF生成模型（NNP-TDGM），将训练集中的术语DEF构造为外显记忆集，在解码器生成（首）义原或关系时，检索与待预测术语概念结构相同或相近的术语所蕴含的核心概念，重要属性和关系类型，辅助模型完成DEF的生成，解决解码器在低频样本上训练不充分的问题。另外，通过提示预训练语言模型获得术语及术语定义内蕴涵概念信息的语义表征向量，改善编码器表征能力不足的问题。经实验验证NNP-TDGM模型生成术语DEF的义原-关系-义原三元组F1值达到31.84%、关系F1值达到53.12%、义原F1值达到51.55%、首义原F1值达到68.53%，相对于基线方法分别提升了3.38%，1.45%，1.08%，0.48%。

**关键词：** Hownet知网；DEF生成；最近邻；提示表征

# NNP-TDGM: Nearest Neighbor Prompt Term DEF Generation Model

**Sijia Shen　Peiyan Wang　Shengren Wang　Libang Wang**
School of Computer Science, Shenyang Aerospace University,
Shenyang, Liaoning 110136, China
Information Center, AVIC Shenyang Aircraft Company Limited,
Shenyang, Liaoning 110003, China
wangpy@sau.edu.cn　wangsr002@avic.com
{shensijia,wanglibang}@stu.sau.edu.cn

## Abstract

To address the issue that the automatic generation of term Definition (DEF) based on Knowledge Database Mark-up Language of HowNet, this paper proposes a Nearest Neighbor Prompt Term DEF Generation Model (NNP-TDGM). This model constructs the term DEF in the training set as explicit memory sets.When generating (first) sememes or roles,the decoder retrieves the core concepts,important attributes,and role types contained in terms with the same or similar conceptual structures to help the model generate DEF. This model addresses the insufficient training problem of the decoder on low-frequency samples. Furthermore, the model improves the problem of insufficient representation ability of the encoder by prompting the pre-trained language model to obtain semantic representation embeddings containing the conceptual information in terms and term definitions. Experimental evidence shows that the sememe-role-sememe triple F1 reaches 31.84%, the role F1 reaches 53.12%, the sememe F1

* 通讯作者
©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版
**基金项目:**全国科技名词审定委员会科研项目(YB2022015)，国家自然科学基金(U1908216)

reaches 51.55% and the first sememe F1 reaches 68.53%. These values show an improvement of 3.38%, 1.45%, 1.08% and 0.48%, respectively, compared to the baseline.

**Keywords:** Hownet , DEF generation , Nearest Neighbor , Prompt representation

## 1 引言

HowNet (董振东et al., 2007)是中文自然语言处理领域使用最为广泛的语义知识库，用于描述汉语和英语词语所代表的概念、以及概念与概念之间、概念的属性与属性之间的关系。HowNet使用义原及义原间关系描述概念，使用知识库描述语言（Knowledge Database Mark-up Language，KDML）作为概念形式化表示方法，形式化表示的概念被称为DEF（Definition）。HowNet被广泛应用于词语表征(Niu et al., 2017)、词义消歧 (Hou et al., 2020)、语言建模 (Gu et al., 2018)、反向字典构建 (Zheng et al., 2020)等任务中。但是HowNet主要面向通用领域，针对专业领域如航空领域的自然语言理解任务其支持能力尚显不足。

张等 (2014)面向航空领域的术语语义知识库构建方法进行了研究，但其采用的是人工或半自动的方法，效率有限。在自动生成义原方面，Xie等 (2017)将词和义原转化为向量，利用协同过滤和矩阵分解方法为新词推荐义原。Jin等 (2018)将字符级内部信息引入汉语词汇义原预测中，提出字符增强语义预测框架，用来缓解由于外部信息的排他性带来的问题。杜等 (2020)利用局部相关性，词语的各个义原和定义文本中的词语之间存在语义匹配关系，提出义原相关池化模型为新词预测义原。Qi等 (2020)构建基于多语言百科全书BabelNet的义原知识库，对多种语言的词语进行义原标注。上述研究将义原预测任务简化为多标签分类任务，未考虑义原的层级结构。Ye等 (2022)提出基于Transformer的义原树生成模型，该方法判断义原与义原之间是否存在关系，没有判断义原与义原间的关系类别，不是完整的词语DEF。吕 (2022)提出基于Seq2Tree (Dong and Lapata, 2016)的DEF自动生成方法，该方法能够生成完整的术语DEF，某种程度上解决了术语DEF自动生成问题。

本文以文献(吕嘉, 2022)方法为基础，进一步研究了术语DEF自动生成问题，提出基于最近邻提示表征的术语DEF生成模型（Nearest Neighbor Prompt Term DEF Generation Model, NNP-TDGM）。主要面向两方面问题：一是由于现有方法以术语为输入，采用“端到端”的编码器-解码器模型生成术语DEF。编码器较难获取术语所蕴含的核心概念及其重要属性，以及核心概念与属性之间，属性与属性之间的复杂关系，进而无法形成准确的语义表征向量。二是由于HowNet有2214种义原与89种关系，在有限训练样本的条件下，义原和关系的训练数据分布不均匀，导致解码器在低频样本下训练不充分。针对上述问题，本文提出两种解决途径：

（1）提出基于提示的术语语义信息表征方法。现有研究表明预训练语言模型蕴涵大量的语义知识 (Zhang et al., 2019)，基于此，本文通过提示预训练语言模型获得术语及术语定义内蕴涵概念信息的语义表征向量，并将其整合到解码器中，改善编码器表征能力不足的问题。

（2）提出基于最近邻的DEF生成方法。将训练集中的术语DEF分解为首义原、义原和关系，首义原代表术语的核心概念，义原细化描述术语的重要属性，关系揭示术语核心概念与属性及属性与属性之间的复杂关系，分别构造首义原外显记忆集，义原外显记忆集和关系外显记忆集。在解码器生成（首）义原或关系时，通过检索与待预测术语概念结构相同或相近的术语中蕴含的核心概念、重要属性和关系类型，辅助模型完成DEF生成，从而弥补低频样本下模型训练不充分问题。

本文在航空术语数据集上进行了实验，NNP-TDGM模型生成术语DEF三元组（义原-关系-义原）$F1$值达到31.84%、关系$F1$值达到53.12%、义原$F1$值达到51.55%、首义原$F1$值达到68.53%，相对于基线方法分别提升了3.38%，1.45%，1.08%，0.48%。实验结果验证了所提出方法的有效性。

本文第2节介绍文本表示与检索增强的相关研究。第3节详细阐述基于最近邻提示表征的术语DEF生成模型，包括整体框架与各模块实现细节。第4节为实验结果与分析，介绍实验数据集，对比模型、评价指标、参数设置、生成DEF实验、消融实验以及对各模块影响分析实验。最后为结论与展望。


(c) 2024 中国中文信息学会计算语言学专业委员会

## 2 相关研究

### 2.1 文本表示

文本表示是将文本转化为连续稠密向量 (曾骏et al., 2024)。预训练语言模型BERT (Devlin et al., 2019)在文本表示方面得到了广泛的应用。Yan等 (2021)和Gao等 (2021)通过对比学习的方法优化预训练语言模型的句表示，同时也缓解了句表示的奇异值快速下降的问题。Su等 (2021)发现传统机器学习中的白化操作可以增强BERT句子表征的各向同性。Jiang等 (2022)通过提示和对比学习相结合的方法提取预训练语言模型的句表示来减小词嵌入带来的偏置。

在生成术语DEF中，可利用的信息仅为术语及术语定义，并且较难构造出大量的训练数据，因此无法使用需要大量训练语料或丰富上下文的文本表示方法。因此，本文提出基于提示的术语语义信息表征方法，提示预训练语言模型获得术语及术语定义内蕴涵概念信息的语义表征向量，能够充分利用预训练语言模型具有的大量语义知识。

### 2.2 检索增强

最近，检索增强方法被应用于自然语言处理的各类任务中。其方法是构建外显记忆模块，在模型预测过程中检索外显记忆，获得实例，通过实例与模型联合预测，改善预测结果，如Khandelwal等 (2020)提出的K近邻机器翻译，Guu等 (2020)提出的K近邻语言模型。Wang等 (2020)在人机对话中检索相似历史对话并结合知识图谱，提升对话的流畅性与信息丰富度。Fan等 (2021)提出通过检索增强为对话生成任务提供先验知识。Liu等 (2023)提出一种用于低资源神经机器翻译的K最近邻迁移学习方法，该方法提高了低资源神经机器翻译的性能。

现有研究主要面向翻译与语言模型等序列生成任务，在树形结构生成任务中还鲜有研究。本文在树形结构的术语DEF生成任务上研究检索增强方法，可为结构化数据生成提供研究借鉴。

## 3 基于最近邻提示表征的术语DEF生成模型

### 3.1 问题定义

本文目标是输入术语$e$，自动生成术语$e$的DEF。举例说明，术语“主燃区”的DEF为“{place|地方:RelateTo={burn|焚烧},modifier={primary|主}}”，其DEF的义原树结构如图1所示。“place|地方”为首义原，代表“主燃区”的核心概念为某个特定位置，而“burn|焚烧”和“primary|主”为义原，分别指明了“主燃区”的具有的主要功能和属性。“RelateTo”和“modifier”为关系，揭示了属性与核心概念存在相关和修饰关系。

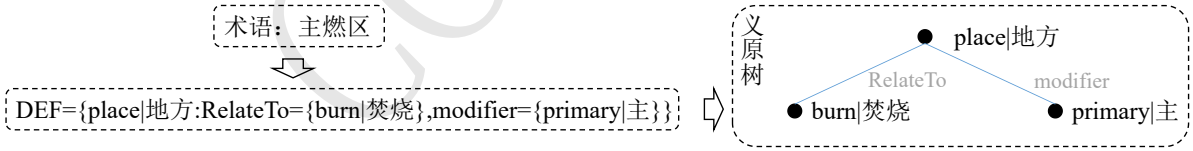



Figure 1: DEF和义原树示例

DEF生成问题可转化为义原树生成问题 (Ye et al., 2022; 吕嘉, 2022)。一个义原树可以分为从根节点到叶子节点的多条路径。假设不同的路径都是独立的，则生成义原树的概率如式(1)$p(T|e)$所示。每个路径的概率如式(1)$p(S|e)$所示。利用马尔可夫假设，进一步将路径分解为父-子路径，基于父节点生成子节点的概率如式(1)$p(s_t|e)$所示。

$$p(T|e)=\prod_{S\in T}p(S|e);\quad p(S|e)=\prod_{t=1}^{N_S}p(s_t|e,S_{0:t-1});\quad p(s_t|e)=\prod_{t=1}^{N_S}p(s_t|e,s_{t-1}); \tag{1}$$

其中，$T$表示的是术语$e$的义原树，$S$表示$T$中的路径，由义原及关系构成，是“义原-关系-义原-关系-...-义原”序列。$N_S$是$S$的长度，$t$为当前时间点，$s_t$是$S$当前的路径状态，对于义原树路径，$s_t$为义原或是关系，$S_{0:t-1}$是从根节点到$S$的$t-1$个时间步的路径。


(c) 2024 中国中文信息学会计算语言学专业委员会

### 3.2 模型总框架

本文提出基于最近邻提示表征的术语DEF生成模型（NNP-TDGM）。本模型沿袭了SRS（吕嘉，2022)模型所用的义原框架填充思想，得到形如“义原-关系-义原”形式的输出。NNP-TDGM主要由基于提示的术语语义信息表征模块、树形解码器、KNN模块、基于最近邻的术语DEF生成模块构成。NNP-TDGM模型结构如图2所示。

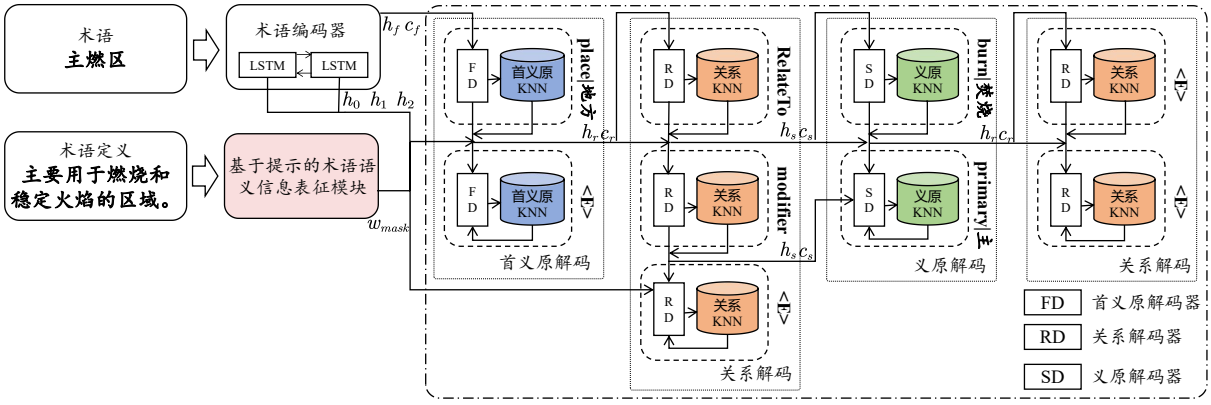



Figure 2: NNP-TDGM模型结构图

### 3.3 基于提示的术语语义信息表征模块

本文提出一种基于提示的术语语义信息表征方法。通过提示从预训练语言模型中获得术语语义的向量化表征，作为补充信息辅助DEF生成。以术语及术语定义为基础，构建3种提示信息：术语、定义、术语和定义。其中，术语定义是对术语的解释说明文本，例如术语“主燃区”的定义是“主要用于燃烧和稳定火焰的区域。”。每种提示信息包括4类模板，分别命名为CLS、MASK、PMASK和LMASK。模板中“[术语]”代表术语，“[定义]”表示术语定义，“[MASK]”表示掩码提示符。每类模板与3类提示信息组合形成12种提示模板，如TCLS表示使用术语信息的CLS类模板。12种提示模板列入表1中。CLS类模板是将提示信息$\{x_{CLS}, x_0, \cdots, x_n, x_{SEP}\}$输入到预训练语言模型中，获取输出层CLS向量作为语义表征向量$\boldsymbol{w}_{CLS}$，如式(2)所示。MASK类模板将输入（术语、定义、术语及定义）与提示符“[MASK]”拼接，并将拼接之后的提示语句$\{x_{CLS}, x_0, \cdots, x_{MASK}, x_{SEP}\}$输入到预训练语言模型中，获取输出层“[MASK]”提示符输出的向量作为语义表征向量$\boldsymbol{w}_{MASK}$，如式(3)所示。PMASK和LMASK类模板都将输入放到更加平滑的语境中，同样添加提示符“[MASK]”，获取预训练语言模型“[MASK]”提示符输出的向量作为语义表征向量$\boldsymbol{w}_{MASK}$，如式(3)所示。

$$\boldsymbol{w}_{CLS}, \boldsymbol{w}_0, \cdots, \boldsymbol{w}_n, \boldsymbol{w}_{SEP} = BERT\left(x_{CLS}, x_0, \cdots, x_n, x_{SEP}\right) \tag{2}$$

$$\boldsymbol{w}_{CLS}, \boldsymbol{w}_0, \cdots, \boldsymbol{w}_{MASK}, \boldsymbol{w}_{SEP} = BERT\left(x_{CLS}, x_0, \cdots, x_{MASK}, x_{SEP}\right) \tag{3}$$

| 模板名称 | 模板 | 模板名称 | 模板 |
|---|---|---|---|
| TCLS | [术语] | DCLS | [定义] |
| TMASK | [术语][MASK] | DMASK | [定义][MASK] |
| TPMASK | 这句话“[术语]”的意思是[MASK] | DPMASK | 这句话“[定义]”的意思是[MASK] |
| TLMASK | “[术语]”这句话的意思是[MASK] | DLMASK | “[定义]”这句话的意思是[MASK] |
| ACLS | [术语][定义] | | |
| AMASK | [术语][定义][MASK] | | |
| APMASK | 这句话“术语:[术语]的定义是:[定义]”的意思是[MASK] | | |
| ALMASK | “术语:[术语]的定义是:[定义]”这句话的意思是[MASK] | | |

Table 1: 12类提示模板

### 3.4 树形解码器

树形解码器包含首义原解码器（First Sememe Decoder，FD）、关系解码器（Role Decoder，RD）和义原解码器（Sememe Decoder，SD）3部分。3个解码器都采用长短期记忆神经网络（LSTM）模型。3个解码器的结构相同，以首义原解码器为例，模型结构如图3中首义原解码器所示。将语义表征向量$\boldsymbol{w}_{MASK}$和解码器当前时刻的隐藏层向量$\boldsymbol{h}_t$以及术语编码器上下文向量$\boldsymbol{g}_t$进行拼接得到拼接向量$\boldsymbol{o}_t$，生成的路径状态概率如式(4)所示。由于树形解码器包含三个解码器，FD得到的概率是$p_{MF}(s_t|e)$，RD得到的概率是$p_{MR}(s_t|e)$，SD得到的概率是$p_{MS}(s_t|e)$。

$$
\begin{aligned}
\boldsymbol{h}_t, \boldsymbol{c}_t &= \text{LSTM}\left(\boldsymbol{h}_{t-1}, \boldsymbol{c}_{t-1}\right) \\
score_i^t &= \frac{\exp\left\langle \boldsymbol{h}_i \cdot \boldsymbol{h}_t \right\rangle}{\sum_{j=1}^{l} \exp\left\langle \boldsymbol{h_j} \cdot \boldsymbol{h}_t \right\rangle} \\
\boldsymbol{g}_t &= \sum_{i=1}^{l} score_i^t \boldsymbol{h}_i \\
\boldsymbol{o}_t &= [\boldsymbol{h}_t; \boldsymbol{g}_t; \boldsymbol{w}_{MASK}] \\
p_M\left(s_t|e\right) &= \text{softmax}\left(W \cdot \boldsymbol{o}_t\right)
\end{aligned}
\tag{4}
$$

其中$\boldsymbol{h}_{t-1}$和$\boldsymbol{c}_{t-1}$分别是$t-1$时刻解码器的隐藏层向量和细胞状态，$\boldsymbol{c}_t$是$t$时刻解码器的细胞状态，$W$是参数矩阵，$\boldsymbol{h}_i$为术语编码器端第i个字符对应的隐藏层向量，$l$为术语$e$的字符数量。

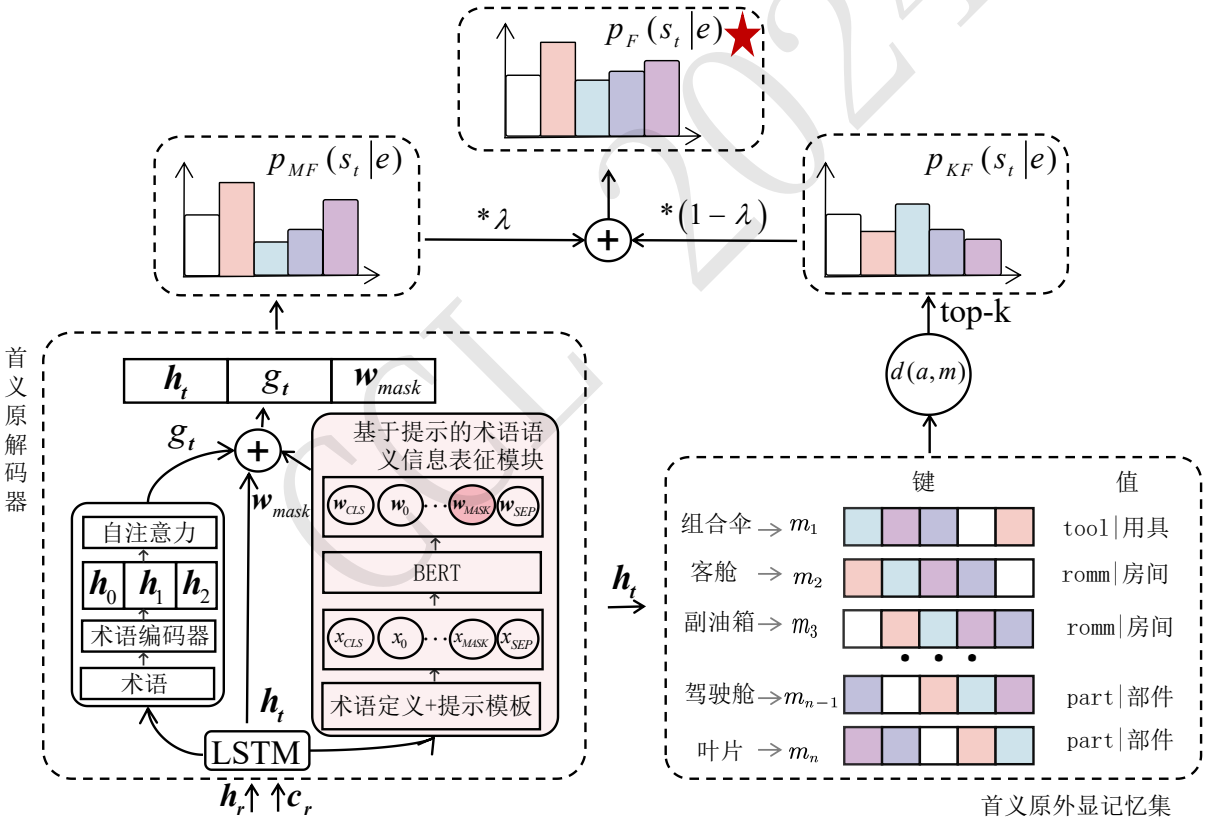



Figure 3: 首义原解码器和首义原KNN联合解码

### 3.5 KNN模块

本文针对3个解码器分别设计3个外显记忆集，分别为首义原外显记忆集，关系外显记忆集和义原外显记忆集。外显记忆集是由键-值$(K,V)$对构成，构建方式如式(6)所示。“键”$K$有三种选择：$t-1$时刻解码器的隐藏层向量$\boldsymbol{h}_{t-1}$，$t$时刻解码器的隐藏层向量$\boldsymbol{h}_t$和$t$时刻解码器的拼接向


(c) 2024 中国中文信息学会计算语言学专业委员会

量$\boldsymbol{o}_t$，如式(5)。首义原外显记忆集和义原外显记忆集的“值”$V$是义原集合$Y_s$，关系外显记忆集的“值”$V$是关系集合$Y_r$。

$$H = \{\boldsymbol{h}_{t-1}, \boldsymbol{h}_t, \boldsymbol{o}_t\} \tag{5}$$

$$(K, V) = \left\{(\boldsymbol{m}, v) \,\middle|\, \boldsymbol{m} \in H, v \in Y_r \text{或} Y_s\right\} \tag{6}$$

在解码时，$\boldsymbol{a} \in H$为查询向量，$\boldsymbol{m}$和$v$为$t$时刻当前解码器的外显记忆集中的键和值。计算$\boldsymbol{a}$与$\boldsymbol{m}$的余弦相似度$d(\boldsymbol{a}, \boldsymbol{m})$，取与$\boldsymbol{a}$余弦相似度最大的$k$个键-值对构成$t$时刻当前解码器的最近邻集合$U_t$，通过式(7)得到$t$时刻KNN概率$p_K(s_t|e)$。由于本文设置三个解码器，所以首义原KNN概率$p_{KF}(s_t|e)$、关系KNN概率$p_{KR}(s_t|e)$、义原KNN概率$p_{KS}(s_t|e)$。

$t$时刻解码器概率$p_M(s_t|e)$和KNN概率$p_K(s_t|e)$加权求和得到$t$时刻路径状态概率$p(s_t|e)$，如式(8)所示，其中$\lambda$是概率权值。

$$p_K(s_t|e) \propto \sum_{(\boldsymbol{m}, v) \in U_t} \mathbb{I}_{s_t = v} \exp(d(\boldsymbol{a}, \boldsymbol{m})) \tag{7}$$

$$p(s_t|e) = \lambda p_M(s_t|e) + (1-\lambda) p_K(s_t|e), \ \lambda \in (0,1) \tag{8}$$

### 3.6 基于最近邻的术语DEF生成模块

DEF具体生成过程按照如下步骤:

**步骤1**：$\boldsymbol{h}_f$和$\boldsymbol{c}_f$是术语编码器输出的最终隐藏状态和最终细胞状态，用于生成术语对应的首义原。

**步骤2**：解码生成首义原。解码生成首义原过程如图3所示。调用首义原解码器，得到首义原解码器概率$p_{MF}(s_t|e)$并调用首义原KNN得到首义原KNN概率$p_{KF}(s_t|e)$将二者加权求和如式(8)获取首义原概率$p_F(s_t|e)$，最终输出概率最大的首义原。有些术语的首义原有多个，比如“故障检测与定位”，该术语的首义原是“decide|决定”和“check|查”。所以每次解码时会得到一个首义原或者是终结符“<E>”，当生成结果是首义原时保存本次输出的隐藏状态$\boldsymbol{h}_r$和细胞状态$\boldsymbol{c}_r$用于生成首义原所对应的关系。

**步骤3**：解码生成关系。将步骤2的$\boldsymbol{h}_r$和$\boldsymbol{c}_r$输入关系解码器和关系KNN中，得到关系的概率$p_R(s_t|e)$，概率最大的值即为义原对应的关系。当生成终结符“<E>”或者达到最大数量后停止。如果解码结果是关系，保存本次输出的隐藏状态$\boldsymbol{h}_s$和细胞状态$\boldsymbol{c}_s$用于生成此关系对应的义原。

**步骤4**：解码生成义原。将步骤3输出的$\boldsymbol{h}_s$和$\boldsymbol{c}_s$输入到义原解码器和义原KNN中，得到义原的概率$p_S(s_t|e)$，输出义原后，保存输出的隐藏状态$\boldsymbol{h}_r$和细胞状态$\boldsymbol{c}_r$用于生成义原对应的关系。

**步骤5**：重复步骤3与4，直到步骤3生成了终结符“<E>”或者生成结果所对应的树形结构已经达到了预设的最大深度。

## 4 实验结果与分析

### 4.1 数据集

本文实验数据集取自航空术语语义知识库 (张桂平et al., 2014)，标注了3864个航空术语，在其中选取能在《中国航空百科词典》 (《中国航空百科词典》编辑部, 2000)查找到完整定义的术语。最终本实验数据集细分为17种航空领域类型，共计3146条术语，包含900种义原，105种关系，术语平均长度4.78字符，术语定义平均长度36.54字符。本文将数据随机分为三部分：训练集（70%）、验证集（10%）和测试集（20%）。本数据集的统计信息如表2所示。

| 数据类型 | 术语数 | 义原数 | 关系数 | 三元组数 |
|---|---|---|---|---|
| 训练集 | 2431 | 11224 | 8765 | 8765 |
| 验证集 | 347 | 1610 | 1260 | 1260 |
| 测试集 | 694 | 3173 | 2471 | 2471 |

Table 2: 航空术语语义知识库数据集统计信息


(c) 2024 中国中文信息学会计算语言学专业委员会

### 4.2 对比模型

本文选择LLaMA-7B、ChatGLM-6B、Seq2Tree、SSR、SRS、TSTG和TaSTG模型作为对比模型:

LLaMA-7B (Touvron et al., 2023)，基于Transformer架构的大型语言模型。ChatGLM-6B (Du et al., 2022)，基于General Language Model（GLM）架构的对话语言模型。Seq2Tree (Dong and Lapata, 2016)，一种变种的Seq2Seq方法。在解码时，将Seq2Seq一轮顺序解码变为多批次的顺序解码，每批次解码出树形结构中一个节点的所有子节点。SSR (吕嘉, 2022)，使用添加注意力机制的树形解码器生成父和子义原以及两个义原之间的关系，得到“义原-义原-关系”三元组形式的输出，最终得到术语对应的DEF。SRS (吕嘉, 2022)，交替使用添加注意力机制的义原解码器和关系解码器，得到“义原-关系-义原”三元组形式的输出，最终得到术语对应的DEF。TSTG (Ye et al., 2022)，专为义原树生成任务设计的Transformer模型。其将任务限定为仅生成义原，不考虑生成义原之间的关系。TaSTG (Ye et al., 2022)，为了更好捕捉数据的结构，设计树注意力，分为语义注意力和位置注意力。语义注意力捕获树中两个节点的语义相似性，位置注意力捕获树中节点的拓扑关系。

### 4.3 评价指标

NNP-TDGM模型在测试集上得到的三元组（义原-关系-义原）、关系、义原和首义原的F1值作为评价指标。F1值越高代表识别效果越好。具体计算公式如式(9)，(10)，(11)所示。

$$F1 = \frac{2 \times P \times R}{P + R} \tag{9}$$

$$P = \frac{\text{生成正确}\,(\text{首})\,\text{义原、关系、三元组数}}{\text{所有生成}\,(\text{首})\,\text{义原、关系、三元组数}} \times 100\% \tag{10}$$

$$R = \frac{\text{生成正确}\,(\text{首})\,\text{义原、关系、三元组数}}{\text{真实样本}\,(\text{首})\,\text{义原、关系、三元组数}} \times 100\% \tag{11}$$

### 4.4 模型参数设置

NNP-TDGM模型的超参数设置如表3所示。

| 超参数 | 取值范围 | 超参数 | 取值范围 |
|---|---|---|---|
| 训练轮次 | 500 | 优化器 | AdamW |
| 训练批量 | 500 | Dropout | 0.5 |
| 字向量维度 | 400 | LSTM层数 | 1 |
| 隐藏层维度 | 400 | 随机种子 | 1 |
| 学习率 | 1e-3 | 注意力头数 | 1 |
| $k$ | 5 | $\lambda$ | (0,1) |

Table 3: NNP-TDGM模型参数设置

### 4.5 术语DEF生成实验

表4展示各模型的实验结果。TSTG和TaSTG仅生成义原，不生成义原之间的关系，所以只有义原$F1$值和首义原$F1$值的实验结果。


(c) 2024 中国中文信息学会计算语言学专业委员会

| 模型 | 三元组F1 | 关系F1 | 义原F1 | 首义原F1 |
|---|---|---|---|---|
| TSTG | - | - | 44.26 | 31.95 |
| TaSTG | - | - | 39.90 | 30.88 |
| ChatGLM-6B | 10.26 | 34.59 | 29.42 | 51.36 |
| LLaMA-7B | 11.15 | 36.51 | 31.67 | 51.29 |
| Seq2Tree | 18.97 | 42.98 | 35.41 | 65.22 |
| SSR | 19.76 | 39.70 | 48.22 | **71.37** |
| SRS | 28.46 | 51.67 | 50.47 | 68.05 |
| **NNP-TDGM(ours)** | **31.84** | **53.12** | **51.55** | 68.53 |

Table 4: 各模型实验结果（%）

由表4可见，NNP-TDGM好于对比模型，在三元组、关系、义原和首义原的生成上都有所提高，与对比方法中DEF生成效果最好的SRS模型相比，三元组、关系、义原和首义原$F1$值分别提高了3.38%，1.45%，1.08%，0.48%。

### 4.6 消融实验

#### 4.6.1 NNP-TDGM消融实验

本节进行消融实验，验证基于提示的术语语义信息表征模块（Prompt）与KNN模块（KNN）的有效性。表5展示消融实验结果，其中"-"表示模型不使用该模块，"√"表示模型使用该模块。

| 模型 | KNN | Prompt | 三元组F1 | 关系F1 | 义原F1 | 首义原F1 |
|---|---|---|---|---|---|---|
| NNP-TDGM | √ | √ | 31.84 | 53.12 | 51.55 | 68.53 |
| NN-TDGM | √ | - | 30.88 | 52.96 | 52.10 | 69.25 |
| P-TDGM | - | √ | 30.39 | 53.31 | 51.52 | 69.10 |
| SRS | - | - | 28.46 | 51.67 | 50.47 | 68.05 |

Table 5: 消融实验结果（%）

表5实验结果表明，基于提示的术语语义信息表征模块和KNN模块的加入能够提升模型DEF生成效果。与SRS模型相比，加入KNN模块使三元组、关系、义原和首义原$F1$值分别提高了2.42%，1.29%，1.63%，1.20%。加入基于提示的术语语义信息表征模块使三元组、关系、义原和首义原$F1$值分别提高了1.93%，1.64%，1.05%，1.05%。结果证明将基于提示的术语语义信息表征模块和KNN模块应用于生成术语DEF的有效性。

#### 4.6.2 KNN模块消融实验

为研究KNN模块对术语DEF生成效果的影响。设计了$K_F$：仅首义原解码器加入KNN,$K_R$：仅关系解码器加入KNN和$K_S$：仅义原解码器加入KNN的实验。采用的模型是P-TDGM。实验效果如表6所示。

| 模型 | 三元组F1 | 关系F1 | 义原F1 | 首义原F1 |
|---|---|---|---|---|
| P-TDGM | 30.39 | 53.31 | 51.52 | 69.10 |
| $+K_F$ | 30.53 | 53.31 | 51.69 | **69.80** |
| $+K_R$ | **30.74** | **53.59** | 51.72 | 69.10 |
| $+K_S$ | 30.66 | 53.31 | **52.00** | 69.10 |

Table 6: 解码器加入KNN的实验结果（%）

在解码器下加入KNN（$K_F$、$K_R$和$K_S$）对各解码器性能产生积极效果。$K_F$对首义原F1值提升明显，同样，$K_S$对义原F1值提高明显，$K_R$对三元组和关系F1值提高明显。因此，解码器


(c) 2024 中国中文信息学会计算语言学专业委员会

引入KNN模块是一种有效方法，可以提高模型性能，而且不同解码器加入KNN能有针对性地提升模型在不同任务下的性能。

### 4.7 提示模板影响实验

为验证3.3节12种提示模板对术语DEF生成的影响，设计了P-TDGM使用各类提示模板的实验。模型对12种不同提示模板进行实验得到12种模型。如P-TDGM$_{DLMASK}$模型是使用表1中DLMASK模板的P-TDGM。实验结果如表7所示，同时列出了SRS结果作为参照。

| 模型 | 三元组F1 | 关系F1 | 义原F1 | 首义原F1 |
|---|---|---|---|---|
| SRS | 28.46 | 51.67 | 50.47 | 68.05 |
| P-TDGM$_{TCLS}$ | 26.61 | 50.83 | 49.54 | 67.09 |
| P-TDGM$_{TMASK}$ | 25.47 | 50.86 | 47.06 | 65.22 |
| P-TDGM$_{TPMASK}$ | 26.80 | 49.93 | 48.47 | 65.80 |
| P-TDGM$_{TLMASK}$ | 27.01 | 50.41 | 48.89 | 65.94 |
| P-TDGM$_{ACLS}$ | 29.09 | 52.06 | 51.40 | 69.35 |
| P-TDGM$_{AMASK}$ | 29.64 | 52.79 | 51.12 | 68.77 |
| P-TDGM$_{APMASK}$ | 30.10 | 52.67 | 51.18 | 67.48 |
| P-TDGM$_{ALMASK}$ | 30.75 | 53.20 | 51.07 | 67.76 |
| P-TDGM$_{DCLS}$ | 28.53 | 53.66 | 51.24 | 70.40 |
| P-TDGM$_{DMASK}$ | 29.00 | 51.87 | 52.02 | 70.97 |
| P-TDGM$_{DPMASK}$ | 29.58 | 52.48 | 51.03 | 69.10 |
| P-TDGM$_{DLMASK}$ | 30.39 | 53.31 | 51.52 | 69.10 |

Table 7: P-TDGM使用不同提示模板的实验结果（%）

由表7可见，采用不同提示模板对术语DEF生成效果具有明显的影响，这与预训练语言模型提示敏感的性质相符，进一步分析数据可发现:

(1)提示中仅使用术语信息的P-TDGM$_{TCLS}$、P-TDGM$_{TMASK}$、P-TDGM$_{TPMASK}$与P-TDGM$_{TLMASK}$模型的$F1$值与SRS相比均有小幅度下降。本文认为其原因是，一方面术语仅为一个词或短语，有效提示信息较少，很难从预训练语言模型中提取出有效的语义表征向量。另一方面，术语作为模型输入，在模型内形成了表征向量，该表征向量与预训练语言模型提示生成的向量不同构，也可能产生负面影响。

(2)相比于基于术语信息的提示，使用具有更丰富信息的术语定义的提示实验效果更好。P-TDGM$_{DCLS}$、P-TDGM$_{DMASK}$、P-TDGM$_{DPMASK}$与P-TDGM$_{DLMASK}$的各项指标皆好于对比方法SRS。这表明在提示中加入更多描述术语含义的信息是有必要的。此外，还发现同时使用术语及定义提示与仅使用定义提示的效果差别不大。另外，PMASK类及LMASK类模板与CLS及MASK类模板相比在三元组上F1值有所提升，表明增加更为平滑语言环境是有效的。

综上，实验表明提示设计的必要性，对于利用提示生成的术语语义表征向量问题，应考虑如何在提示语句中加入描述术语含义的信息，并且设计平滑语境使得提示语句更符合语言规律。

### 4.8 检索向量影响实验

在KNN模块中，3个外显记忆集是由键-值对构成，如式(6)所示。每个外显记忆集中的键可使用3种向量：（1）$\boldsymbol{h}_F$、$\boldsymbol{h}_R$与$\boldsymbol{h}_S$分别代表输入到首义原解码器、关系解码器与义原解码器的隐藏层向量，这些向量携带了截至当前时间步之前的所有信息。（2）$\boldsymbol{h}_{FT}$、$\boldsymbol{h}_{RT}$与$\boldsymbol{h}_{ST}$分别代表首义原解码器、关系解码器与义原解码器的隐藏状态，这些向量携带了解码器输出的首义原、关系或义原的信息。（3）$\boldsymbol{o}_{FT}$、$\boldsymbol{o}_{RT}$与$\boldsymbol{o}_{ST}$分别代表首义原解码器、关系解码器与义原解码器的拼接向量。将这3类向量分别作为外显记忆集中的键，研究其对术语DEF生成效果的影响。实验结果如表8所示，同时将SRS模型作为参照，列出了SRS模型的实验结果。表中FD表示首义原解码器、RD表示关系解码器、SD表示义原解码器。


(c) 2024 中国中文信息学会计算语言学专业委员会

| 解码器 | 模型 | 三元组F1 | 关系F1 | 义原F1 | 首义原F1 |
|---|---|---|---|---|---|
| | SRS | 28.46 | 51.67 | 50.47 | 68.05 |
| FD | $+\boldsymbol{h}_F$ | 28.51 | 51.67 | 50.48 | 68.10 |
| | $+\boldsymbol{h}_{FT}$ | 28.51 | 51.67 | 50.41 | 67.95 |
| | $+\boldsymbol{o}_{FT}$ | 28.51 | 51.67 | 50.68 | 69.10 |
| RD | $+\boldsymbol{h}_R$ | 29.57 | 52.57 | 51.16 | 68.05 |
| | $+\boldsymbol{h}_{RT}$ | 29.33 | 52.43 | 50.73 | 68.05 |
| | $+\boldsymbol{o}_{RT}$ | 29.49 | 52.81 | 51.05 | 68.05 |
| SD | $+\boldsymbol{h}_S$ | 29.04 | 51.65 | 50.86 | 68.05 |
| | $+\boldsymbol{h}_{ST}$ | 29.13 | 51.65 | 51.00 | 68.05 |
| | $+\boldsymbol{o}_{ST}$ | 29.17 | 51.65 | 50.86 | 68.05 |

Table 8: $\lambda$=0.8时不同解码器选择不同向量的实验结果（%）

实验结果表明，对于各解码器采用不同向量均能提高模型性能。综合多方面拼接向量$\boldsymbol{o}$的效果更好。在首义原解码器中使用$\boldsymbol{o}_{FT}$效果最佳，首义原F1值提升1.05%。在关系解码器中使用$\boldsymbol{o}_{RT}$对关系$F1$值提升最高，提升了1.14%。在义原解码器中使用$\boldsymbol{h}_{ST}$使义原$F1$值提升0.53%。

### 4.9 $\lambda$值影响实验

本节研究KNN中概率权值$\lambda$对NNP-TDGM模型效果的影响。$\lambda$在区间$(0,1)$内取值，NNP-TDGM模型各$F1$值如图4所示。从图4中可以观察到，$F1$值随着$\lambda$的增加呈现上升趋势，在达到峰值后有所回落。表明$\lambda$对$F1$值是产生影响的。$\lambda$小于0.5时，KNN概率占比较大，实验结果低于基准值，表明KNN仅能提供辅助作用，不能作为主导。当$\lambda = 0.6$时三元组$F1$值最高，并且随着$\lambda$的增加，三元组$F1$值变化趋于平稳。

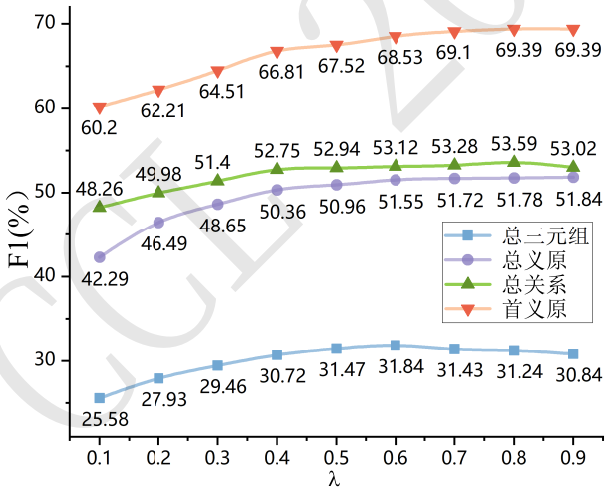



Figure 4: $k = 5$时NNP-TDGM模型各$F1$值随$\lambda$变化曲线

### 4.10 实例分析

本节对KNN模块的作用进行了分析，附录中展示了术语、正确的DEF、SRS和NNP-TDGM生成的DEF，阐述了首义原KNN、关系KNN和义原KNN的检索情况。

从检索实例来看，首义原KNN能检索出与待预测术语概念结构相同或相近的术语中蕴含的核心概念。如实例1中，预测术语“壁温试验”的首义原时，首义原KNN检索到概念结构相近的术语“高温试验”和“耐久性试验”，并且得到这些术语对应的核心概念“fact|事情”。义原KNN能检索出与待预测术语概念结构相同或相近的术语的重要属性。如实例2中，预测术语“自动驾驶仪”的义原时，义原KNN能检索出术语“积分式自动驾驶仪”和“驾驶仪”的重要属性“drive|驾驶”。关系KNN能检索出与待预测术语概念结构相同或相近的术语的关系类型。如实例3中，


(c) 2024 中国中文信息学会计算语言学专业委员会

预测术语“超临界机翼”的关系时，关系KNN检索出术语“有限翼展机翼”和“层流机翼”的关系类型“whole”。

结果表明，KNN通过检索与待预测术语概念结构相同或相近的术语中蕴含的核心概念、重要属性和关系类型，辅助模型完成DEF的生成，从而弥补低频样本下模型训练不充分的问题。

## 5 结论与展望

本文提出一种基于最近邻提示表征的术语DEF生成模型（NNP-TDGM）。提出基于提示的术语语义信息表征方法，基于术语、定义、术语和定义3类信息构建4类共12种模板，通过提示预训练语言模型获得术语及术语定义内蕴含概念信息的语义表征向量，将其整合到解码器中。提出基于最近邻的DEF生成方法，将训练集中的术语DEF构造成3种外显记忆集。在解码器生成（首）义原或关系时，通过检索与待预测术语概念结构相同或相近的术语中蕴含的核心概念、重要属性和关系类型，辅助模型完成DEF生成。通过实验验证了基于提示的术语语义信息表征方法及基于最近邻的DEF生成方法能够改善术语DEF生成效果。对于利用提示生成的术语语义表征向量问题，应考虑如何在提示语句中加入描述术语含义的信息，同时应设计平滑语境使得提示语句更符合语言规律。对于首义原、关系和义原解码器分别加入KNN方法皆有效果。

在未来工作中，将考虑如何动态选取KNN的参数，减少噪声的引入，充分发挥KNN检索性能以及如何获得并加入更多术语信息，例如含该术语的文本片段，相关其他术语等。

## 参考文献

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Li Dong and Mirella Lapata. 2016. Language to logical form with neural attention. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*, pages 33–43.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Angela Fan, Claire Gardent, Chloé Braud, and Antoine Bordes. 2021. Augmenting transformers with knn-based composite memory for dialog. *Transactions of the Association for Computational Linguistics*, 9:82–99.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6894–6910.

Yihong Gu, Jun Yan, Hao Zhu, Zhiyuan Liu, Ruobing Xie, Maosong Sun, Fen Lin, and Leyu Lin. 2018. Language modeling with sparse product of sememe experts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4642–4651.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, pages 3929–3938.

Bairu Hou, Fanchao Qi, Yuan Zang, Xurui Zhang, Zhiyuan Liu, and Maosong Sun. 2020. Try to substitute: An unsupervised chinese word sense disambiguation method based on hownet. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 1752–1757.


(c) 2024 中国中文信息学会计算语言学专业委员会

Ting Jiang, Jian Jiao, Shaohan Huang, Zihan Zhang, Deqing Wang, Fuzhen Zhuang, Furu Wei, Haizhen Huang, Denvy Deng, and Qi Zhang. 2022. Promptbert: Improving BERT sentence embeddings with prompts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 8826–8837.

Huiming Jin, Hao Zhu, Zhiyuan Liu, Ruobing Xie, Maosong Sun, Fen Lin, and Leyu Lin. 2018. Incorporating chinese characters of words for lexical sememe prediction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2439–2449.

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2020. Nearest neighbor machine translation. In *Proceedings of the 9th International Conference on Learning Representations, ICLR 2021,Virtual Event, Austria, May 3-7, 2021*.

Shudong Liu, Xuebo Liu, Derek F. Wong, Zhaocong Li, Wenxiang Jiao, Lidia S. Chao, and Min Zhang. 2023. knn-tl: k-nearest-neighbor transfer learning for low-resource neural machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1878–1891.

Yilin Niu, Ruobing Xie, Zhiyuan Liu, and Maosong Sun. 2017. Improved word representation learning with sememes. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 2049–2058.

Fanchao Qi, Liang Chang, Maosong Sun, Sicong Ouyang, and Zhiyuan Liu. 2020. Towards building a multilingual sememe knowledge base: Predicting sememes for babelnet synsets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8624–8631.

Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. 2021. Whitening sentence representations for better semantics and faster retrieval. *arXiv preprint arXiv:2103.15316*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Jian Wang, Junhao Liu, Wei Bi, Xiaojiang Liu, Kejing He, Ruifeng Xu, and Min Yang. 2020. Improving knowledge-aware dialogue generation via knowledge base question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9169–9176.

Ruobing Xie, Xingchi Yuan, Zhiyuan Liu, and Maosong Sun. 2017. Lexical sememe prediction via word embeddings and matrix factorization. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 4200–4206.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. Consert: A contrastive framework for self-supervised sentence representation transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 5065–5075.

Yining Ye, Fanchao Qi, Zhiyuan Liu, and Maosong Sun. 2022. Going "deeper": Structured sememe prediction via transformer with tree attention. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 128–138.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: enhanced language representation with informative entities. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1441–1451.

Lei Zheng, Fanchao Qi, Zhiyuan Liu, Yasheng Wang, Qun Liu, and Maosong Sun. 2020. Multi-channel reverse dictionary model. In *Proceedings of the AAAI conference on artificial intelligence*, pages 312–319.

《中国航空百科词典》编辑部. 2000. 中国航空百科词典. 航空工业出版社.

吕嘉. 2022. 面向复合型术语的知网DEF自动生成方法研究. 硕士学位论文, 沈阳航空航天大学.


(c) 2024 中国中文信息学会计算语言学专业委员会

张桂平, 刁丽娜, and 王裴岩. 2014. 基于HowNet的航空术语语义知识库的构建. 中文信息学报, 28:92–101.

曾骏, 王子威, 于扬, and 文俊浩. 2024. 自然语言处理领域中的词嵌入方法综述. 计算机科学与探索, 18:24–43.

杜家驹, 岂凡超, 孙茂松, and 刘知远. 2020. 基于局部语义相关性的定义文本义原预测. 中文信息学报, 34:1–9.

董振东, 董强, and 郝长伶. 2007. 知网的理论发现. 中文信息学报, 4:3–9.

## 附录.实例分析

| 术语 | 壁温试验 | | | | |
|---|---|---|---|---|---|
| 正确 DEF | {fact\|事情:CoEvent={experiment\|实验:content={Temperature\|温度:host={part\|部件:PartPosition={skin\|皮},whole={building\|建筑物}}},means={try\|尝试}}} | | | | |
| SRS | {**experiment\|实验**:CoEvent={experiment\|实验:content={Temperature\|温度:host={physical\|物质}}}} | | | | |
| NNP-TDGM | {**fact\|事情**:CoEvent={experiment\|实验:content={Temperature\|温度},means={try\|尝试:host={physical\|物质}}}} | | | | |
| 首义原 | $p_{SRS}$ | $p_{FKNN}$ | 检索出的核心概念 | 检索出的核心概念对应的术语和 DEF | |
| | | | | 术语 | DEF |
| fact\|事情 | 0.45 | 1.0 | fact\|事情 | 高温试验 | {**fact\|事情**:CoEvent={experiment\|实验:PatientAttribute={Temperature\|温度:host={physical\|物质},modifier={GreaterThanNormal\|高于正常}}}} |
| | | | fact\|事情 | 耐久性试验 | {**fact\|事情**:CoEvent={experiment\|实验:content={Quality\|质量:host={physical\|物质}},means={try\|尝试}}} |

Figure 5: 实例分析1

预测术语“壁温试验”的首义原时，首义原KNN检索到与“壁温试验”概念结构相近的术语“高温试验”和“耐久性试验”，并且得到这些术语对应的核心概念“fact|事情”。

| 术语 | 自动驾驶仪 | | | | |
|---|---|---|---|---|---|
| 正确 DEF | {tool\|用具:#instrument={drive\|驾驭:modifier={automatic\|自动}}} | | | | |
| SRS | {tool\|用具:#instrument={**automatic\|自动**},modifier={automatic\|自动:modifier={automatic\|自动}}} | | | | |
| NNP-TDGM | {tool\|用具:#instrument={**drive\|驾驭**:modifier={automatic\|自动}}} | | | | |
| 义原 | $p_{SRS}$ | $p_{SKNN}$ | 检索出的重要属性 | 检索出的重要属性对应的术语和 DEF | |
| | | | | 术语 | DEF |
| drive\|驾驭 | 0.59 | 0.66 | drive\|驾驭 | 积分式自动驾驶仪 | {tool\|用具:#instrument={**drive\|驾驭**:manner={automatic\|自动}},RelateTo={Pattern\|样式:host={knowledge\|知识:domain={math\|数学}}}} |
| | | | drive\|驾驭 | 驾驶仪 | {tool\|用具:#instrument={**drive\|驾驭**}} |

Figure 6: 实例分析2

预测术语“自动驾驶仪”的义原时，义原KNN能检索到与“自动驾驶仪”概念结构相近的术语“积分式自动驾驶仪”和“驾驶仪”，并且得到这些术语对应的重要属性“drive|驾驶”。


(c) 2024 中国中文信息学会计算语言学专业委员会

<table>
<tr><td>术语</td><td colspan="5">超临界机翼</td></tr>
<tr><td>正确 DEF</td><td colspan="5">{part|部件:PartPosition={wing|翅},whole={aircraft|飞行器},modifier={special|特别}}</td></tr>
<tr><td>SRS</td><td colspan="5">{part|部件:PartPosition={wing|翅},modifier={aircraft|飞行器},whole={aircraft|飞行器}}</td></tr>
<tr><td>NNP-TDGM</td><td colspan="5">{part|部件:PartPosition={wing|翅},whole={aircraft|飞行器},modifier={special|特别}}</td></tr>
<tr><td rowspan="2">关系</td><td rowspan="2">$p_{SRS}$</td><td rowspan="2">$p_{RKNN}$</td><td rowspan="2">检索出的关系类型</td><td colspan="2">检索出的关系类型对应的术语和 DEF</td></tr>
<tr><td>术语</td><td>DEF</td></tr>
<tr><td rowspan="2">whole</td><td rowspan="2">0.45</td><td rowspan="2">1.0</td><td>whole</td><td>有限翼展机翼</td><td>{part|部件:PartPosition={wing|翅},whole={aircraft|飞行器},modifier={few|少},RelateTo={Length|长度:host={part|部件:PartPosition={wing|翅},whole={aircraft|飞行器}}}}</td></tr>
<tr><td>whole</td><td>层流机翼</td><td>{part|部件:PartPosition={wing|翅},whole={aircraft|飞行器},modifier={polished|光:degree={most|最}}}</td></tr>
</table>

Figure 7: 实例分析3

预测术语“超临界机翼”的关系时，关系KNN能检索到与“超临界机翼”概念结构相近的术语“有限翼展机翼”和“层流机翼”，并且得到这些术语对应的关系类型“whole”。

CCL 2024

# SpanCS：面向跨语言代码生成的片段级语码转换

朱庆福[1], 周士祺[1], 王硕[2], 张致铭[1], 王昊钰[3], 陈麒光[1], 车万翔[1*]
[1]哈尔滨工业大学
[2]清华大学
[3]北京邮电大学
{qfzhu, sqzhou, zmzhang, qgchen, wxche}@ir.hit.edu.cn
wangshuo.thu@gmail.com
wanghaoyu666@bupt.edu.cn

## 摘要

跨语言代码生成旨在将英语到代码的生成能力迁移至其他自然语言。翻译-训练（Translate-Train）和语码转换（Code-Switching）是实现跨语言迁移的两类经典数据增广方法，两者优势互补但尚未有效结合。为此，本文提出了一种面向跨语言代码生成的片段级语码转换（SpanCS）方法。首先，该方法利用语码转换框架关联源语言上下文与目标语言片段，以促进多种语言的交互和对齐。其次，该方法利用翻译-训练方法从完整的源语言翻译中提取目标语言片段，以保证增广数据与原始数据间的语义一致性。为了公平地评价多种自然语言之间代码生成的性能差异，本文通过人工翻译与校验，基于HumanEval构建了包含10种自然语言的多语言代码生成评测基准MHumanEval。该基准上的三个主干模型的实验结果表明，SpanCS在跨语言代码生成任务上一致优于前人的数据增广方法。

**关键词：** 跨语言；代码生成；语码转换

# SpanCS: Span-Level Code-Switching for Cross-Lingual Code Generation

**Qingfu Zhu[1], Shiqi Zhou[1], Shuo Wang[2], Zhiming Zhang[1], Haoyu Wang[3], Qiguang Chen[1], Wanxiang Che[1*]**
[1]Harbin Institute of Technology
[2]Tsinghua University
[3]Beijing University of Posts and Telecommunications
{qfzhu, sqzhou, zmzhang, qgchen, wxche}@ir.hit.edu.cn
wangshuo.thu@gmail.com
wanghaoyu666@bupt.edu.cn

## Abstract

Cross-lingual code generation aims to transfer the ability of generating code from English to other natural languages (NLs). Translate-train and Code-switching are two common data augmentation (DA) approaches for cross-lingual transfer, which complement each other but have not been effectively combined. To this end, we propose a span-level code-switching (SpanCS) approach for cross-lingual code generation. First, it leverages the code-switching framework to correlate source language context and target language span to model the interaction and alignment among multiple languages. Second, it utilizes the translate-train approach to extract target language span from a complete source language translation, ensuring the semantic consistency between augmented data and original data. To fairly evaluate the discrepancy of code generation across multiple NLs, we construct MHumanEval, a multilingual code generation

* 通讯作者
©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

benchmark that includes 10 NLs, based on HumanEval via manual translation and verification. Experiments on the benchmark across three backbones show that SpanCS consistently outperforms conventional DA approaches for cross-lingual code generation.

**Keywords:** cross-lingual , code generation , code-switching

## 1 引言

代码生成（Code Generation）任务旨在根据给定的自然语言描述生成相应的编程语言代码 (Chen et al., 2021; Li et al., 2022)。近年来，代码大模型技术蓬勃发展，带动了更多种类的编程语言数据向大模型的汇集 (Li et al., 2023; Roziere et al., 2023; Nijkamp et al., 2023; Daya Guo, 2024)，进而推动了代码生成任务由单编程语言（如Python）向多编程语言的扩展 (Feng et al., 2020; Zheng et al., 2023b)。与此同时，由于全球95%人口的母语为非英语的自然语言 (Guo, 2018)，进一步将代码生成任务扩展至多自然语言同样至关重要。然而，由于缺乏非英语到代码的平行训练数据与评测基准，目前多自然语言代码生成的研究仍处于起步阶段。

跨语言微调可以有效缓解多语言任务缺少带标签[1]的目标语言数据的问题，其微调带标签的源语言数据，利用大模型的语言迁移能力，将微调源语言获得的能力迁移至目标语言 (Zheng et al., 2021)。将跨语言微调应用于代码生成任务，即跨语言代码生成[2]，即可通过微调英语到代码数据实现多自然语言到代码的生成。在此基础上，将带标签的源语言（通常为英语）微调数据通过数据增广拓展至目标语言，可以进一步提升跨语言迁移的效果。常见的数据增广方法包括翻译-训练（Translate-Train） (Singh et al., 2019; Chai et al., 2022)以及语码转换（Code-Switching） (Liu et al., 2020; Qin et al., 2021)。

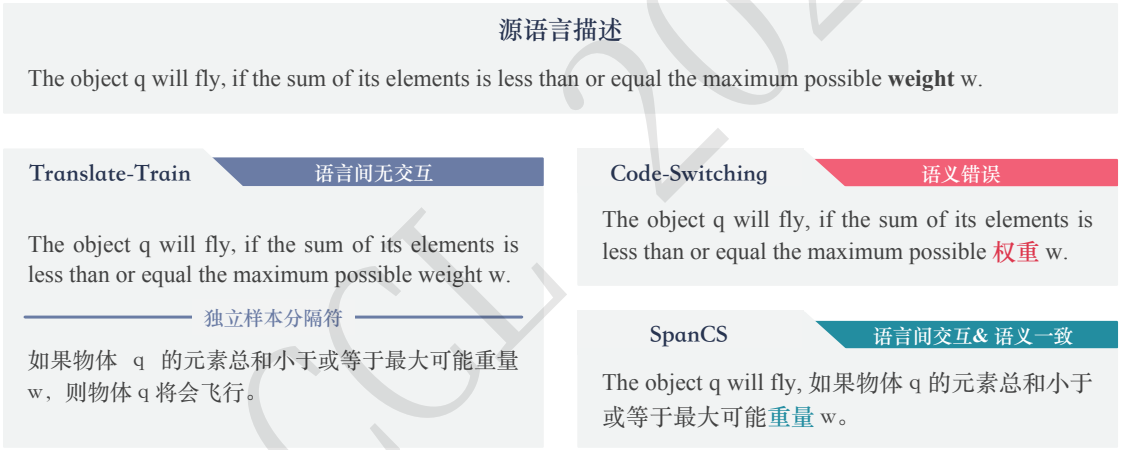



Figure 1: 翻译-训练（Translate-Train）、语码转换（Code-Switching）、片段级语码转换（SpanCS）方法示例（为避免歧义下文使用英文名称指代具体模型，中文名称泛指技术范式）。SpanCS中的“重量”为该上下文环境下源语言英语单词“weight”的正确中文翻译，表示质量大小。Code-Switching中的“权重”为错误中文翻译，表示重要程度。

具体而言，翻译-训练方法通过将带标签数据中的源语言翻译为目标语言，实现对目标语言的数据增广（如图 1所示，简洁起见，图中仅展示了各增广方法对自然语言的处理，省略了所有方法的标签，即对应的代码）。然而，翻译-训练方法中源语言样本与目标语言样本在训练过程中独立学习，缺少不同语言间的交互与对齐，因此限制了源语言到目标语言的迁移效率 (Zheng et al., 2021)。相对地，语码转换方法将源语言中的单词随机替换为目标语言中的翻译，从而有效地在语言模型建模过程中关联了不同语言。例如语码转换样本“除了1和该数本身以外不再有其他的因数的数被称为prime number”，根据中文关于素数的解释预测英文概念“prime number”的学习过程，显式地将两者在语义层面进行对齐，进而促进了两种语言的对

[1]对于代码生成任务而言，标签即为与自然语言描述对应的代码。
[2]本文中，“跨语言”和“多语言”的语言指**自然语言**，而非编程语言。


(c) 2024 中国中文信息学会计算语言学专业委员会

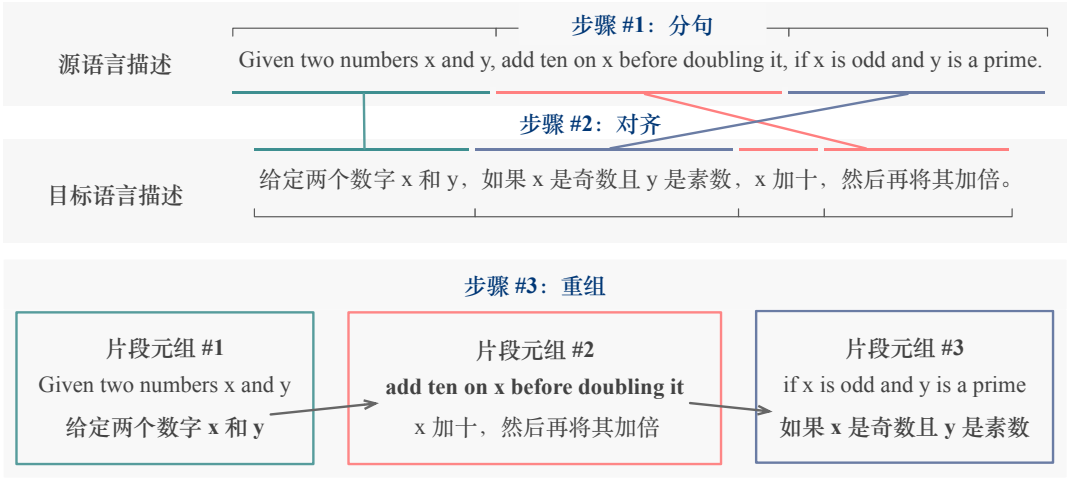



Figure 2: SpanCS的三个步骤的示例。**分句**：将源语言描述翻译成目标语言描述，然后将两者分割成子句。**对齐**：将两种语言的子句按片段对齐，其中，片段由一系列连续的子句构成。**重组**：从每个对齐的片段元组中选择一个源语言片段或目标语言片段，重组成语码转换描述。

齐。然而，当被替换的源语言单词存在多个翻译时，错误的翻译将与源语言上下文产生语义冲突（如图 1所示）。幸运的是，翻译-训练方法可以在翻译过程中通过考虑上下文获取准确的目标语言翻译，因此，基于翻译-训练增强语码转换可以自然地融合两类方法的优势，强化源语言与目标语言交互的同时保持语义一致性，构建高效高质量的增广数据。

为了融合翻译-训练与语码转换的优势，本文提出了一种片段级语码转换（**Span**-level **C**ode-**S**witching, SpanCS）方法，基于源语言的目标语言翻译选择性地替换源语言中的片段。如图 2所示，SpanCS首先将<源语言描述-代码>数据中的源语言描述翻译为目标语言描述，并将两者分割成子句。其次，以片段（存在语义等价匹配的最短连续子句）为单位对齐源语言与目标语言的子句。最后，在代码大模型的指导下，按序从每个对齐的<源语言片段-目标语言片段>元组中选择其一，组合成语码转换的增广样本。

直观上，SpanCS不仅在单个样本内显式地关联了源语言与目标语言，有效建模了语言间的交互与对齐关系，而且以上下文作为参考信息完成源语言片段到目标语言片段的转换，保证了转换前后的语义一致性。此外，为了使语码转换过程可以更好地适配下游代码生成任务，重组步骤在代码大模型的指导下展开，意味着SpanCS可以提供面向具体任务的定制化接口。

跨语言代码生成的效果可以通过多语言评测基准进行评价。然而，包含多种自然语言描述的现有代码生成评测基准，如MCoNaLa (Wang et al., 2023a)和ODEX (Wang et al., 2022)，语言之间并非完全对齐，即某种语言的一个样本在其他语言上可能不存在语义等价的平行样本。因此，一种语言的评测结果优于另一种语言，既可能是由于模型处理两种语言的能力不同，也可能是由于两种语言的评测样本难度不同，进而导致难以公平地评价模型在不同语言上的性能差异。为此，本文将HumanEval (Chen et al., 2021)的英语描述扩展至九种其他自然语言，构建了一个名为MHumanEval的新评测基准。

本文的贡献总结如下：

- 本文结合翻译-训练与语码转换的优势，提出了一种新的跨语言数据增广方法SpanCS，可以将英语代码生成能力迁移至多种自然语言。
- 本文构建了多语言代码生成评测基准MHumanEval，涵盖共10种自然语言（包括英语），用于公平地评价模型跨语言代码生成能力。
- 在三种主干代码大模型上的实验结果表明，SpanCS一致优于前人的跨语言数据增广方法。

## 2 相关工作

### 2.1 跨语言代码生成

在缺少带标签多语言数据的条件下，面向多种自然语言进行代码生成的关键是迁移源语言


(c) 2024 中国中文信息学会计算语言学专业委员会

（通常是英语）上习得的能力至其他语言。语码转换将源语言描述的部分单词替换为目标语言的翻译，提供了一种数据高效的解决方案 (Qin et al., 2021; Liu et al., 2020)。其数据高效特性主要体现在以下两方面，首先，将源语言描述替换为语码转换描述不必引入额外的训练样本（相对地，翻译-训练方法为每个样本引入了一条新的目标语言训练样本）；其次，语码转换的单个样本内包含多种语言，通过语言模型显式建模了语言间的交互，更高效地实现了语言间的知识迁移（相对地，翻译-训练方法中语义等价的不同语言的样本被视为独立样本分别单独训练，语言间无交互关系，知识迁移效率较低）。语码转换中的单词替换通常借助双语词典完成，因此当源语言单词在目标语言中存在多个翻译时，替换过程将存在较为明显的不确定性。为解决该问题，可以训练一个转写模型根据源语言描述直接生成其语码转换后的版本 (Tarunesh et al., 2021; Gautam et al., 2021)。本文提出的SpanCS与传统的语码转换方法的不同之处在于，它可以参考上下文信息解决词典替换的不确定性问题，并且不需要英语到语码转换的平行训练数据来学习转写模型。值得注意的是，翻译-训练与语码转换两类方法均存在曝光偏置，即训练格式不完全等价于推理格式。该现象存在的主要原因如下：一方面，训练过程引入其他语言，尤其是富资源语言（如英语），可以借助大模型的迁移学习能力提升目标语言的性能。另一方面，考虑到实际应用中为每一种语言单独开发部署一个专用模型成本较高，因此训练一个多语言模型更符合实际应用场景。

SpanCS的另一个相关工作是ERNIE-Code (Chai et al., 2022)，其在预训练过程中引入了大量的平行机器翻译数据，以构建多语言代码大模型。SpanCS与之不同之处在于，ERNIE-Code是一个**预训练**基础模型，多语言代码生成能力源自海量的<英语-代码>形式的代码生成数据与<英语-非英语>形式的机器翻译数据，两类数据以英文为枢轴实现隐式对齐。SpanCS是一种跨语言**微调**方法，旨在增广出<语码转换-代码>形式的样本，显式构造非英语到代码的对齐数据，以期通过微调进一步提升代码大模型已有的多语言代码生成能力。

### 2.2 多语言代码生成评测基准

一个涵盖多种语言的全面的评测基准对于公平评价多语言能力至关重要。在代码生成领域，Wang et al. (2023b)通过持续收集和标注西班牙语、日语和俄语的<自然语言-代码>平行数据，扩展了以英语为中心的CoNaLa评测基准 (Yin et al., 2018)。在此基础上，Wang et al. (2022)进一步补充了单元测试用例，使得评价可以从功能正确性的角度展开，更符合人类在软件开发过程中的验证代码质量的真实评价过程。

上述基准与本文提出的MHumanEval的区别在于：首先，MHumanEval在不同语言之间是平行的，即不同语言中的自然语言描述在语义上完全等价，由此可以在控制内容难度相同的条件下分析由语言差异导致的性能变化。其次，作为HumanEval (Chen et al., 2021)的一种扩展，MHumanEval原则上可以方便地桥接其他基于HumanEval的扩展工作。例如，将多自然语言的MHumanEval与多编程语言的HumanEval-X组合，即可构建出多自然语言到多编程语言的更为全面的多语言代码生成基准。

## 3 方法

本节将详细介绍SpanCS，一种用于跨语言代码生成任务的数据增广方法。如图 2所示，首先SpanCS将源语言描述翻译为目标语言描述，并分别将两者切分为子句（§3.2）。然后，将源语言描述和目标语言描述在片段级别上对齐，生成一系列<源语言片段-目标语言片段>形式的元组。此处的片段被定义为源/目标语言描述中的最短连续子句，且在目标/源语言描述中存在语义等价的对应片段（§3.3）。最后，通过在每个对齐的元组中选择一个源语言片段或目标语言片段进行重组，即可构建一个新的语码转换描述（§3.4）。

### 3.1 符号标记

本文使用大写字母表示随机变量（例如，$V$），使用小写字母表示随机变量的索引（例如$i$），使用粗体字母表示向量（例如：$\boldsymbol{V} = \{V_1, ..., V_n\}$）。以此，源语言描述和目标语言描述分别表示为$\boldsymbol{S} = \{S_1, ..., S_m\}$和$\boldsymbol{T} = \{T_1, ..., T_n\}$，其中$S_i$和$T_i$分别表示它们的第$i$个子句。[3] 片段的起止位置由子句的索引表示，例如，$\boldsymbol{S}$中从第$i$个子句开始到第$j$个子句结束的片段（包含起止索引）表示为$\boldsymbol{S}_{i:j}$。

[3]在分句步骤之前子句的划分是不可见的。


(c) 2024 中国中文信息学会计算语言学专业委员会

Translate the text below in {Chinese}, be careful not to translate variable names, function names, code, database schema, graph and HTML.Your translation should not include a prompt, such as: "This is my translation" or "Translation as follows." Here is text:
{Given two numbers x and y, add ten on x before doubling it, if x is odd and y is a prime.}

Figure 3: 将源语言描述翻译为目标语言描述的提示和输入示例。红色和绿色字体分别表示目标语言和待翻译的源语言描述样本。

### 3.2 分句

给定源语言描述$\boldsymbol{S}$，SpanCS将其作为一个整体翻译为目标语言描述$\boldsymbol{T}$以确保$\boldsymbol{S}$和$\boldsymbol{T}$之间的语义一致性。通常，描述中可能包含若干程序相关的成分，例如函数名和单元测试用例等，这些成分在翻译过程中应保持不变，否则将破坏后续生成的代码的正确性。例如，函数名的更改将破坏生成的函数与其调用程序之间预定义的接口。然而，机器翻译模型很难区分程序成分与纯文本。为此，本文精心设计了相应的指令，提示GPT-3.5 Turbo完成这项任务，具体指令信息如图 3所示。

在此之后，SpanCS将源语言描述和目标语言描述分别划分为子句，划分过程可以形式化表示为$\text{SEG}(\boldsymbol{S}) = \{S_1, ..., S_m\}$，$\text{SEG}(\boldsymbol{T}) = \{T_1, ..., T_n\}$，其中SEG表示分句函数。具体而言，分句的粒度应尽可能地细致，以支持后续精准的片段对齐和多样的片段级语码转换。为实现该目标，本文尽可能多地收集了常用的分隔符，并通过启发式正则实现子句划分。

### 3.3 对齐

语码转换中被替换的源语言成分应当与替换后的目标语言成分语义对齐。然而，由于翻译后的目标语言可能具有与源语言不同的子句数量和顺序，因此对齐难以在子句级别进行。如图2所示，源语言的三个子句被翻译成目标语言的四个子句，且源语言的第二个子句同时对应目标语言的第三和第四个子句。为应对上述问题，SpanCS的对齐将在片段（Span）级别进行。对齐过程可形式化表示为将$\{S_1, ..., S_m\}$和$\{T_1, ..., T_n\}$划分为一系列对齐的片段元组$\{(\boldsymbol{S}_{1:i}, \boldsymbol{T}_{1:p}), ..., (\boldsymbol{S}_{j:m}, \boldsymbol{T}_{q:n})\}$，其中$\boldsymbol{S}_{i:j}$和$\boldsymbol{T}_{p:q}$是语义等价的最短连续子句序列。

具体而言，SpanCS采用贪心算法实现片段对齐以确保对齐效率。[4] 首先，选择$\boldsymbol{S}$和$\boldsymbol{T}$的第一个子句作为一个候选片段元组: $\boldsymbol{C} = (\boldsymbol{S}_{1:1}, \boldsymbol{T}_{1:1})$. 其次，将$\boldsymbol{S}$和$\boldsymbol{T}$的第二个子句分别扩展到当前候选片段元组中，以获得两个包含更多子句的新候选$\boldsymbol{C}_s = (\boldsymbol{S}_{1:2}, \boldsymbol{T}_{1:1})$和$\boldsymbol{C}_t = (\boldsymbol{S}_{1:1}, \boldsymbol{T}_{1:2})$。然后，通过LaBSE度量 (Feng et al., 2022)计算每个候选内源语言片段与目标语言片段之间语义相似度。如果$\boldsymbol{C}$具有最高相似度，则表示长度最短的候选$(\boldsymbol{S}_{1:1}, \boldsymbol{T}_{1:1})$已经在语义上对齐，因此可以被接收为一个正式的片段元组。否则，$\boldsymbol{C}_s$和$\boldsymbol{C}_t$之间具有较高相似度的元组将作为新的$\boldsymbol{C}$，并重复上述扩展和比较过程，直到扩展额外的子句不能进一步提高相似度为止。

在上述对齐算法中，部分$\boldsymbol{S}$或$\boldsymbol{T}$末尾的子句可能无法成功对齐，因为将它们扩展到最后一个片段元组后不能带来进一步的相似度提升。尽管如此，SpanCS仍然将其附加到最后一个片段元组，以确保所有子句都存在对齐对象，以便在下一步中进行语码转换。

### 3.4 重组

给定对齐后的片段元组构成的序列，重组可以形式化表示为：按顺序从每个元组中选择源语言片段或目标语言片段，将所有选中片段拼接构成语码转换描述。因此，对于包含$L$个片段元组的样本，搜索空间将为$2^L$，通过枚举所有可能全局搜索最优解的方案时间复杂度过高，无法达到实际应用的需求。为了缓解上述问题，本文确定了最优语码转换描述的两个关键特征，以启发式查找一个近似解。首先，最优语码转换描述应当是连贯的，即每个子句与该子句的上下文保持连贯。其次，作为服务于特定任务（如代码生成）的中间步骤，最优语码转换描述应同时是任务相关的，即通过合理地选择和组织语码，辅助提升下游具体任务的性能。

**基于困惑度的重组** 为了将上述两个因素引入搜索过程，本文设计了一种基于困惑度的（perplexity, PPL）的重组机制。它采用束搜索（beam search）算法在代码大模型的指导下

[4]本文同时尝试了更复杂的动态规划算法SentAlign (Steingrimsson et al., 2023)，详见§5.5。


(c) 2024 中国中文信息学会计算语言学专业委员会

查找具有最小困惑度的不同语言片段的组合。在搜索的每一步中，分别将下一个片段元组的源语言片段和目标语片段扩展到束中的所有候选上，然后保留按困惑度排序的top-$K$候选，其中$K$表示束的大小。

直观上，大规模数据上训练的大模型所计算困惑度可以有效地评价连贯性。同时，代码大模型的引入使得重组过程可以更为精准地识别和处理程序片段。该方法的一个潜在问题是困惑度指标可能更倾向于单语描述而非语码转换描述。但基于以下两点考虑，本文并未对该倾向性进行干预。首先，从语码转换结果上看，约42%的样本产生了语码转换，因此该倾向性并未对语码转换的出现产生严重的影响。其次，本文认为该倾向性可以一定程度上降低语义错误的语码转换出现的概率。该倾向性的存在，使得仅当目标语言片段相对于源语言片段在困惑度上存在优势、且该优势足以抵消单一语言的倾向性时，才可以进行语码转换，因此可以屏蔽掉不确定性大的转换，仅保留高质量的转换。

**基于先验的加权** 在基于困惑度的重组的基础上，本文进一步引入了基于先验的加权机制来平衡特定语言的倾向。实际上，由于各语言的token数量在词表中的占比是不均衡的，大模型计算的困惑度在不同语言之间存在偏差。为消除该偏差，本文以源语言为枢轴对所有目标语言的困惑度进行了加权。具体而言，首先从源语言和目标语言训练数据中随机抽取100个样本，分别计算它们的困惑度，其中源语言结果记为$P_s$，目标语言结果记为$P_t$。然后，计算$P_s/P_t$作为每个目标语言片段在束搜索过程中的先验，以此平衡大模型在语言之间的倾向性。

## 4 多语言代码生成基准MHumanEval

平行的多语言评测基准对于定量对比各种语言的性能至关重要。然而，目前代码生成领域尚未有这样的基准被提出，严重阻碍了该领域的发展。为此，本文将HumanEval (Chen et al., 2021)的英语（en）描述扩展到其他九种语言构建了MHumanEval。参考 Qin et al. (2023)的选择，本文的多语言包括孟加拉语（bn）、德语（de）、西班牙语（es）、法语（fr）、日语（ja）、俄语（ru）、斯瓦希里语（sw）和泰卢固语（te）以及汉语（zh）。

具体而言，本文采用专业的商业服务[5]完成MHumanEval源语言描述到目标语言描述的翻译。与常规翻译任务不同，代码生成任务的描述中包含一些程序片段，例如函数名等，这些片段在翻译中应保持不变。为此，本文在翻译之前对该类片段进行了人工标注。在翻译过程中，每一个样本先后由一名译员翻译和一位编辑审校。为了保证构建数据的质量，所有选定的译员和编辑均具备专业认证资质（例如德语、西班牙语、泰语专业八级，日语国际一级等）和三年以上的工作经验，累计翻译量超过200万字。成本方面，每个源语言（英语）单词翻译成目标语言的平均成本为0.065美元。

## 5 实验

### 5.1 数据

SpanCS将代码生成微调数据中的自然语言描述增广到多种其他自然语言。本文实验中的微调数据为evol-codealpaca-v1，[6] 它是evol-instruct (Luo et al., 2023)的一个开源复现，在Code Alpaca (Chaudhary, 2023)的基础上借助自指令技术扩展而来。完整的数据集包含111,272个样本，增广至$N$种目标语言后，对于翻译-训练等方法而言微调样本将同步扩展$N$倍。为了进行公平的比较和控制微调成本，本文随机选取了10,000个样本展开实验，并将发布该选定数据集以供后续研究参考和对比。

### 5.2 基线方法

本文将SpanCS和零样本以及跨语言微调两种范式下的基线进行了比较。具体基线如下：

- **Base**：以零样本的方式直接提示基础大模型根据目标语言描述生成代码。
- **Cross-Lingual**：在<源语言描述-代码>微调数据上训练，利用大模型的迁移能力提升目标语言上能力。

[5]https://f.youdao.com/
[6]https://huggingface.co/datasets/theblackcat102/evol-codealpaca-v1


(c) 2024 中国中文信息学会计算语言学专业委员会

- **Translate-Train**：将<源语言描述-代码>数据翻译成<目标语言描述-代码>数据，并将两者混合用于微调的跨语言数据增广方法 (Hu et al., 2020)。公平起见，该方法的目标语言描述使用了SpanCS分句步骤中的GPT-3.5 Turbo的翻译结果（详见§3.2）。

- **Code-Switching**：另一种跨语言数据增广方法，以一定的概率（本文中为0.7）随机选取源语言描述中的单词替换为目标语言翻译。

值得注意的是，本文的基线方法中未列入基于转写的语码转换方法 (Tarunesh et al., 2021; Gautam et al., 2021)，因为用于构建该类方法的语码转换训练数据对于大多数目标语言而言在现实中很难收集。

### 5.3 训练与推理设置

所有基线方法和SpanCS都是模型无关的，因此可以适用于各种代码大模型。为了提供全面的效果对比，本文选取了代码生成领域应用最广泛的三种代码大模型作为主干模型开展实验，包括：CodeGen2.5-7B-mono (Nijkamp et al., 2023)、CodeLlama-7B-python (Roziere et al., 2023)和DeepSeekCoder-7B-base (Daya Guo, 2024)。简洁起见，默认情况下，下文将省略代表具体参数与版本的后缀。

| 主干模型 | 方法 | bn | de | en | es | fr | ja | ru | sw | te | zh | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **CodeGen** | Base | 6.7 | 9.1 | 10.9 | 10.3 | 11.5 | 10.9 | 9.1 | 9.7 | 7.9 | 11.5 | 9.8 |
| | Cross | 24.3 | 37.1 | 46.3 | 45.7 | 39.6 | 36.5 | **42.0** | 27.4 | 24.3 | 34.1 | 35.7 |
| | Trans | **28.0** | **40.2** | 45.1 | 43.2 | 40.8 | 35.9 | 38.4 | **29.2** | 23.1 | **42.0** | 36.6 |
| | CS | **28.0** | 37.1 | **47.5** | 45.7 | 37.1 | 38.4 | 40.2 | 26.2 | **26.8** | 36.5 | 36.4 |
| | SpanCS | **28.0** | 39.0 | 46.9 | **46.3** | **45.7** | **42.0** | 41.4 | 26.8 | 25.6 | 38.4 | **38.0** |
| **CodeLlama** | Base | 23.7 | 31.1 | 42.6 | 34.1 | 34.1 | 33.5 | 35.9 | 26.2 | 21.9 | 34.7 | 31.8 |
| | Cross | 30.4 | 43.9 | 50.0 | **54.2** | 47.5 | 39.6 | 41.4 | **31.7** | **29.2** | **45.1** | 41.3 |
| | Trans | **34.7** | 42.0 | 50.6 | 50.6 | 50.0 | 38.4 | 46.3 | 26.2 | 25.6 | 43.2 | 40.8 |
| | CS | 31.7 | **47.5** | 50.6 | 51.2 | 46.3 | 40.2 | 46.3 | 26.2 | **29.2** | 42.6 | 41.2 |
| | SpanCS | 34.1 | 42.0 | **56.0** | 53.6 | **53.0** | **42.0** | **46.9** | 30.4 | **29.2** | **45.1** | **43.2** |
| **DeepSeek** | Base | 20.1 | 34.7 | 44.5 | 34.7 | 32.3 | 42.0 | 39.0 | 29.2 | 23.1 | 47.5 | 34.7 |
| | Cross | 39.6 | 61.5 | 62.8 | 62.8 | **67.0** | **57.9** | 59.7 | 32.9 | 34.7 | 59.7 | 53.9 |
| | Trans | **40.8** | **63.4** | 64.0 | 62.1 | 65.8 | 54.8 | 59.1 | **42.0** | 32.9 | 59.7 | 54.5 |
| | CS | 38.4 | **63.4** | 60.9 | 61.5 | 66.4 | 57.3 | 59.7 | 39.6 | **37.1** | 57.9 | 54.2 |
| | SpanCS | **40.8** | 61.5 | **67.0** | **66.4** | 65.8 | **57.9** | **66.4** | 39.6 | 32.9 | **62.8** | **56.1** |

Table 1: Cross-Lingual (Cross), Translate-Train (Trans), Code-Switching (CS)和SpanCS在三个代码主干大模型（7B）下的pass@1结果。最后一列(Avg.) 表示九种语言的平均性能。

实验中三个主干模型下的所有方法（基线方法和SpanCS）均采用了统一的训练程序。参考目前监督微调范式下的SOTA方法evol-instruct (Luo et al., 2023)的设置，本文的batch大小设置为512，最大长度设置为2048。所有参数由AdamW (Loshchilov and Hutter, 2017)优化器进行更新，其中学习率为1e-3，warmup步数为1,000，训练过程持续三个epoch。推理过程通过贪心策略解码输出代码，并通过pass@1 (Chen et al., 2021)指标对结果进行评价。在SpanCS方法中，基于困惑度的重组的束的大小设置为16，源语言为英语，目标语言为MHumanEval中的其他九种自然语言。

### 5.4 实验结果

各方法在MHumanEval评测基准上的pass@1结果如表1所示。三个主干模型的性能排名从高到低依次为DeepSeekCoder、CodeLlama、CodeGen2.5，大体上呈现出与各主干模型训练数据量相一致的趋势。在每个主干模型中，Base方法由于缺少针对指令跟随能力以及跨语言能力的微调训练，因此其效果显著弱于其他方法。Cross-Lingual方法在<英语-代码>数据上进行了微调，补齐了指令遵循能力，因此相比于Base方法有了极为显著的提升，但跨语言能力的学习仅限于预训练阶段，仍有较大的提升空间。Code-Switching与Cross-Lingual相比效果相当或略有优势，证明了在单个样本中通过语码转换建模语言间交互关系的有效性，该范式同时也是SpanCS所基于的范式。Translate-Train的稳定性相对


(c) 2024 中国中文信息学会计算语言学专业委员会

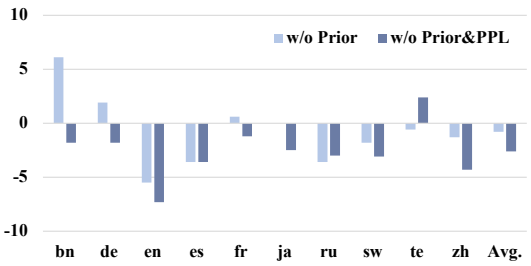



Figure 4: SpanCS去除基于先验的加权(w/o Prior) 与去除整体基于困惑度的重组(w/o Prior&PPL)的消解实验。每个柱状图代表消解前与消解后的pass@1差值。

You are an impartial rater and need to rate the pairs. Below are segmented sentences from two different languages and the pairings of these small sentences. The paired sentences are required to be semantically aligned as much as possible. The pairing can be many-to-many, but the fine-grainedness needs to be as high as possible. The score range is 0-10. The higher the score, the better the alignment of the paired sentences.The formate of score is [[rate]],for example:'Score:[[10]]'
{Given two ⋯}; {给定两个⋯}; {<Given two numbers x and y, 给定两个数字 x 和 y>, ⋯}

Figure 5: 基于GPT-4的对齐策略评价的提示词。待评价的样本按红色部分所示的格式输入，包含：完整的源语言描述、目标语言描述以及对齐后的片段元组。

较差，其在DeepSeekCoder和CodeLlama主干模型上优于Cross-Lingual和Code-Switching，但在CodeGen2.5主干模型上弱于其他方法。本文认为这是由于不同主干模型的多语言能力存在差异，在低资源语言场景下（例如sw与te），Cross-Lingual与Translate-Train在CodeLlama上的差异显著大于在CodeGen2.5、DeepSeekCoder上的差异。

相比之下，SpanCS的平均表现一致优于所有基线方法，且该提升具备统计显著性（t-test, $p$ <0.05）。一方面，SpanCS进一步提升了Code-Switching的性能，证明了借助翻译-训练方法，以片段为单位进行语码转换保持语义一致的有效性。另一方面，相比于Translate-Train，SpanCS降低了CodeLlama主干模型在低资源语言（sw与te）场景下的不稳定性，表明SpanCS在跨语言迁移中更为有效。同时，上述迁移是通过语码转换范式实现的，该范式显式地在单个样本内建模了不同语言之间的交互。综上，SpanCS同时利用了翻译-训练和语码转换的优势，高质量地保证了增广前后的语义一致性、高效地建模了语言间的交互对齐关系，从而提高了整体性能。

### 5.5 实验分析

为了进一步了解SpanCS的有效性，本文以性能最好的DeepSeekCoder为测试平台进行了更为深入的分析。

**消解实验** 为了研究基于先验的加权和基于困惑度的重组对SpanCS整体效果的贡献，本文对相应模块进行了消解，图4展示了消解前与消解后的性能差值。首先，仅从基于困惑度的重组中去除基于先验的加权，结果记为w/o Prior。其次，将基于困惑度的重组（包含基于先验的加权）整体替换为一种随机策略，即随机选取一个源语言描述的片段替换为与其对齐的目标语言片段，结果记为w/o PPL。实验结果显示，消解前后的性能差值大多为负，表明这两个机制对最终性能均存在积极的贡献。其中，te与其他语言表现不一致的主要原因在于，大模型对少资源te语言处理能力较差。作为一个参考，表1中三个主干模型下五种方法的平均性能，te的结果是所有语言中的最低值。因此，基于困惑度的重组模块利用大模型计算te的困惑度时，可能存在更大的误差。

将源语言和目标语言的连续子句对齐成片段元组是SpanCS的关键步骤，为验证默认的贪心对齐策略的有效性，本文从九种目标语言中各抽取500个样本进行了评价。对齐效果的评价是一个较难的任务，待评价样本同时涉及多种语言与编程知识，因此人工评价需要精通各语言的编程人员才能胜任，现实中很难匹配到合适的标注者。因此本文参考前人工作(Zheng et


(c) 2024 中国中文信息学会计算语言学专业委员会

al., 2023a)，使用GPT-4开展评价。该工作表明，GPT-4在各类开放性任务的评价上具有与人类可比的表现与一致性。图5展示了用于评价的提示词，输入的每个待评价样本包含三个字段，分别为完整的源语言描述、目标语言描述以及对齐的片段元组，输出为0-10等级的评分。评价结果如表2所示，贪心策略和基于动态规划的SentAlign策略 (Steingrimsson et al., 2023)均获得了高于9.5分以上的评分，证明两者均有良好的对齐效果。相比之下，默认的贪心策略片段对齐效果（表2，Greedy vs. SentAlign）以及基于对齐结果的下游代码生成（表4，SpanCS vs. SpanCS+SentAlign）均略低于SentAlign，但对齐速度是SentAlign的4.7倍。在计算资源充足的条件下，基于SentAlign实现SpanCS的片段对齐可以进一步提升代码生成的性能。

| 方法 | bn | de | es | fr | ja | ru | sw | te | zh | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|
| Greedy | 9.44 | 9.58 | 9.82 | 9.74 | 9.43 | 9.76 | 9.59 | 8.81 | 9.65 | 9.54 |
| SentAlign | 9.51 | 9.76 | 9.88 | 9.77 | 9.57 | 9.83 | 9.67 | 9.14 | 9.75 | **9.65** |

Table 2: SpanCS对齐步骤中基于贪心的策略（默认）与基于SentAlign动态规划的策略对齐效果。每种对齐策略的分数由GPT-4按0-10等级进行评分。

| 方法 | en | es | ja | ru | Avg. |
|---|---|---|---|---|---|
| Cross-Lingual | 50.3 | 44.4 | **44.5** | **59.1** | 49.6 |
| Translate-Train | 47.6 | 44.4 | 40.8 | 54.3 | 46.8 |
| Code-Switching | 50.3 | 45.5 | 40.8 | 58.3 | 48.7 |
| SpanCS | **50.8** | **50.0** | 41.4 | 57.1 | **49.8** |

Table 3: 各方法在ODEX评测基准上的pass@1结果。

| 方法 | bn | de | en | es | fr | ja | ru | sw | te | zh | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Cross-Lingual | 26.8 | 34.1 | 43.2 | 37.8 | 42.0 | **34.1** | 35.9 | 24.3 | 25.6 | 40.8 | 34.5 |
| Translate-Train | **28.0** | 34.1 | 42.0 | 38.4 | 39.0 | **34.1** | **38.4** | 26.8 | **26.8** | 38.4 | 34.6 |
| Code-Switching | 25.0 | 32.3 | 40.8 | 38.4 | 38.4 | 32.9 | 37.8 | 25.6 | **26.8** | 35.9 | 33.4 |
| SpanCS | 25.6 | 34.7 | **47.5** | **42.6** | 42.6 | 32.3 | 37.1 | **28.6** | 23.7 | 39.0 | 35.4 |
| SpanCS+SentAlign | **28.0** | **39.6** | 45.7 | 41.4 | **43.9** | 32.3 | 37.1 | 28.0 | 25.6 | **42.0** | **36.4** |

Table 4: 各方法在DeepSeekCoder-1.3B-Base主干模型下的pass@1结果。

**泛化性** 为验证SpanCS的泛化性，表3展示各方法在ODEX评测基准上的结果。Translate-Train的平均性能显著弱于其他方法，主要是因为其在日语（ja）和俄语（ru）中的表现较差，这与MHumanEval基准上的结果相一致。相比之下，SpanCS获得了最好的平均性能，并显著提高了西班牙语（es）的性能，表明其可以更好地将在源语言中学习的知识迁移到其他语言。为进一步验证SpanCS在不同量级模型上的性能，本文进一步对比了各方法在DeepSeekCoder-1.3B-Base[7]上的结果。如表4所示，SpanCS在1B量级的模型上获得了最优平均性能，证明其在模型量级维度上同样具备泛化性。

**训练效率** 图6展示了SpanCS和Translate-Train（DeepSeekCoder上除了SpanCS之外的最佳方法）的pass@1指标随着训练过程的变化。SpanCS以1.3倍的速度达到了Translate-Train的最佳性能，同时，在相同训练量下性能超过Translate-Train 3%，证明其具有更优的训练效率。

值得注意的是，由于英语在所有语言中处于枢轴位置，因此每种语言在消解后未解决的困惑度偏差和翻译错误（如图8中SpanCS-Filtered）会级联作用于英语，导致英语降幅最为明显。尽管如此，作为枢轴的中心位置同时也为英语提供了更多的机会来整合其他语言的改进。

**可视化** 跨语言研究的核心是不同语言相互对齐，以此将源语言知识迁移至目标语言。为了直观地理解对齐的效果，本文在图7中通过PCA (Wold et al., 1987)方法展示了对各种语言的隐层状态的可视化效果。参照Qin et al. (2021)的可视化设置，本文从MHumanEval中随机选取

[7]除此处分析实验外，其余实验均在7B量级模型上完成，模型细节详见§5.3。


(c) 2024 中国中文信息学会计算语言学专业委员会

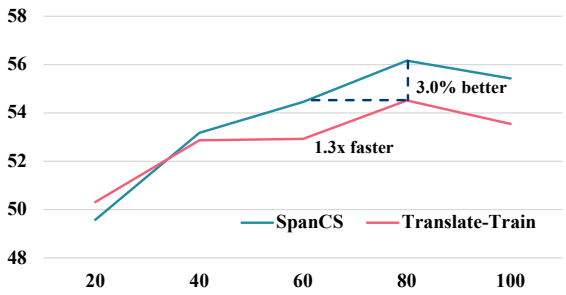



Figure 6: Translate-Train与SpanCS训练效率。横、纵坐标分别表示训练步数与pass@1结果。

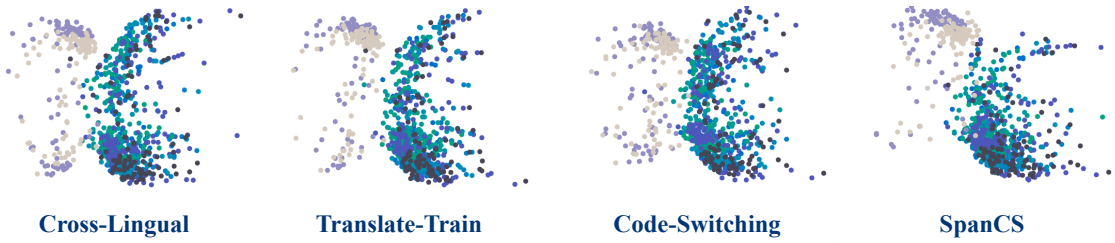



Figure 7: 基于PCA的隐层状态可视化。不同语言由颜色区分。基线方法中的各种语言相对松散地分布在隐层空间的四个角落，SpanCS更为集中地分布在隐层空间的对角线上，语言间的对齐效果更优。

| | 案例 #1 | 案例 #2 |
|---|---|---|
| 源语言描述 | Referring to the dictionary labeled as 'B', ..., It should **be** able to surgically expunge those key-value pairs wherein the values are marked less than 10. // Dictionary B<br>B = {'**one**': 1, '**ten**': 10, '**twenty**': 20, '**zero**': 0} | Implement a Python program to parse the given XML data … Adding **to** it, the XML data includes … <subject>**Maths**</subject>… The program should also handle cases where a subject list may be empty |
| Code-Switching | Referring -కి ఈ 字典 étiqueté as 'B', ..., It উচিৎ **అవ్వు** способный -కు chirurgicalement తీసివేయు those key-value জোড়া wherein ఈ valeurs are marqué কম than 10. // Dictionary B<br>B = {'one': 1, 'ten': 10, 'twenty': 20, 'zero': 0} | Implement ఓ Python программа -కి parse করে donnés XML data … Adding **kufika** it, el XML 資料 includes … <subject>Maths</subject>… The программа উচিৎ কেও కడియం casos где ఉ тема kuorodesha সোমবার kuwa empty. |
| SpanCS-Filtered | Referring to the dictionary labeled as 'B', ..., Inapaswa kuondoa kirafiki key-value ambapo thamani zimehakikiwa kuwa chini ya 10. // Kamusi B B = {'**moja**': 1, '**kumi**': 10, '**ishirini**': 20, '**sifuri**': 0} | একটি পাইথন প্রোগ্রাম বিন্যাস করুন যাতে দেওয়া এক্সএমএল তথ্যটি … এর আরও অংশগুলি সংযোজিত হচ্ছে একটি … <subject>**গণিত**</subject>… The program should also handle cases where a subject list may be empty |
| SpanCS | Kwa kurejelea kamusi iliyoandikwa kama 'B', ..., Inapaswa kuondoa kirafiki key-value ambapo thamani zimehakikiwa kuwa chini ya 10. // Dictionary B<br>B = {'one': 1, 'ten': 10, 'twenty': 20, 'zero': 0} | একটি পাইথন প্রোগ্রাম বিন্যাস করুন যাতে দেওয়া এক্সএমএল তথ্যটি … এর আরও অংশগুলি সংযোজিত হচ্ছে একটি … <subject>Maths</subject>… The program should also handle cases where a subject list may be empty |

Figure 8: Code-Switching, SpanCS, 以及SpanCS在重组步骤中过滤掉的样本示例。红色字体表示Code-Switching中语义错误的替换，例如案例#1中英语短语“be able to”的“be”被替换为了泰卢固语中代表“成为”的单词，案例#2中英语短语“adding to”的“to”被替换为了斯瓦西里语中代表“到达”的单词。绿色字体代表应当保留但被误翻译的程序成分，例如，案例#1中Python词典中的键（key），案例#2中XML数据标签内部的内容。

了100个样本，针对模型中间层（16/32）的隐层状态平均值进行可视化。图中的每一个点代表

CCL 2024

一种语言的一个样本，通过颜色对语言进行区分。

在可视化结果中，更好的对齐效果对应着更为集中的分布，即不同语言互相对齐时，彼此在隐层空间的重叠程度应更高。Cross-Lingual方法在未使用任何显式对齐信息的情况下，不同语言的表示分布在隐层空间的四个角落。Translate-Train的源语言描述和目标语言描述是独立训练的，可视为一种隐式对齐，因此对齐能力有限，对齐效果与Cross-Lingual相比并无显著差异。Code-Switching与其他基线方法之间的区别同样并不显著。其分布边缘甚至略微呈现出发散的趋势。本文认为这是由于其存在语义不一致的替换错误，使得对齐过程产生了混淆。相比之下，SpanCS的各种语言更集中地分布在隐层空间的对角线上，证明了其更优的对齐能力。

**案例分析** 图8展示了Code-Switching、SpanCS以及SpanCS中在基于的困惑度的重组步骤中被过滤掉的样本（记为SpanCS-Filtered）。在缺乏上下文信息的情况下，Code-Switching的替换很容易发生错误。例如，在案例#1中，短语“be able to”的“able”被泰卢固语中代表“成为”的翻译所替换。Translate-Train和SpanCS方法中基于大模型的翻译过程，即便显式地设置了相应的指令要求，应该保留的程序片段仍存在着被翻译的风险，例如案例#1和#2中SpanCS-Filtered的词典的键和XML数据。尽管如此，SpanCS可以在后续基于困惑度的重组阶段借助代码大模型的编程知识屏蔽掉上述存在误翻译现象的候选。

## 6 结论

本文提出了一种面向跨语言代码生成的片段级语码转换方法SpanCS。与现有方法相比，SpanCS有效地将目标语言片段与其源语言上下文相关联，同时保持了两者之间的语义一致性。为了公平地评价多种语言之间代码生成的性能差异，本文构建了一个涵盖共10种语言的MHumanEval评测基准。在三个主干模型上的实验表明，SpanCS的平均性能显著优于所有基线方法。此外，SpanCS更有效地促进了源语言上习得的能力向其他语言的迁移，同时具备良好的训练效率。

## 致谢

本文研究受到国家自然科学基金(No. 62236004，No. 62206078)等项目资助。

## 参考文献

Yekun Chai, Shuohuan Wang, Chao Pang, Yu Sun, Hao Tian, and Hua Wu. 2022. Ernie-code: Beyond english-centric cross-lingual pretraining for programming languages. *arXiv preprint arXiv:2212.06742*.

Sahil Chaudhary. 2023. Code alpaca: An instruction-following llama model for code generation. https://github.com/sahil280114/codealpaca.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harrison Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code.

Dejian Yang Zhenda Xie Kai Dong Wentao Zhang Guanting Chen Xiao Bi Y. Wu Y.K. Li Fuli Luo Yingfei Xiong Wenfeng Liang Daya Guo, Qihao Zhu. 2024. Deepseek-coder: When the large language model meets programming – the rise of code intelligence.

Zhangyin Feng, Daya Guo, Duyu Tang, Nan Duan, Xiaocheng Feng, Ming Gong, Linjun Shou, Bing Qin, Ting Liu, Daxin Jiang, and Ming Zhou. 2020. CodeBERT: A pre-trained model for programming and natural languages. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1536–1547, Online, November. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland, May. Association for Computational Linguistics.


(c) 2024 中国中文信息学会计算语言学专业委员会

Devansh Gautam, Prashant Kodali, Kshitij Gupta, Anmol Goel, Manish Shrivastava, and Ponnurangam Kumaraguru. 2021. CoMeT: Towards code-mixed translation using parallel monolingual sentences. In Thamar Solorio, Shuguang Chen, Alan W. Black, Mona Diab, Sunayana Sitaram, Victor Soto, Emre Yilmaz, and Anirudh Srinivasan, editors, *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 47–55, Online, June. Association for Computational Linguistics.

Philip J Guo. 2018. Non-native english speakers learning computer programming: Barriers, desires, and design opportunities. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–14.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: a massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In *Proceedings of the 37th International Conference on Machine Learning*, pages 4411–4421.

Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, et al. 2022. Competition-level code generation with alphacode. *arXiv preprint arXiv:2203.07814*.

Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, et al. 2023. Starcoder: may the source be with you! *arXiv preprint arXiv:2305.06161*.

Zihan Liu, Genta Indra Winata, Zhaojiang Lin, Peng Xu, and Pascale Fung. 2020. Attention-informed mixed-language training for zero-shot cross-lingual task-oriented dialogue systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8433–8440.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. 2023. Wizardcoder: Empowering code large language models with evol-instruct. *arXiv preprint arXiv:2306.08568*.

Erik Nijkamp, Hiroaki Hayashi, Caiming Xiong, Silvio Savarese, and Yingbo Zhou. 2023. Codegen2: Lessons for training llms on programming and natural languages. *ICLR*.

Libo Qin, Minheng Ni, Yue Zhang, and Wanxiang Che. 2021. Cosda-ml: multi-lingual code-switching data augmentation for zero-shot cross-lingual nlp. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3853–3860.

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2695–2709, Singapore, December. Association for Computational Linguistics.

Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, et al. 2023. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*.

Jasdeep Singh, Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2019. Xlda: Cross-lingual data augmentation for natural language inference and question answering. *arXiv preprint arXiv:1905.11471*.

Steinthor Steingrimsson, Hrafn Loftsson, and Andy Way. 2023. SentAlign: Accurate and scalable sentence alignment. In Yansong Feng and Els Lefever, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 256–263, Singapore, December. Association for Computational Linguistics.

Ishan Tarunesh, Syamantak Kumar, and Preethi Jyothi. 2021. From machine translation to code-switching: Generating high-quality code-switched text. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3154–3169, Online, August. Association for Computational Linguistics.

CCL 2024

Zhiruo Wang, Shuyan Zhou, Daniel Fried, and Graham Neubig. 2022. Execution-based evaluation for open-domain code generation. *arXiv preprint arXiv:2212.10481*.

Zhiruo Wang, Grace Cuenca, Shuyan Zhou, Frank F. Xu, and Graham Neubig. 2023a. MCoNaLa: A benchmark for code generation from multiple natural languages. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 265–273, Dubrovnik, Croatia, May. Association for Computational Linguistics.

Zhiruo Wang, Grace Cuenca, Shuyan Zhou, Frank F. Xu, and Graham Neubig. 2023b. MCoNaLa: A benchmark for code generation from multiple natural languages. In Andreas Vlachos and Isabelle Augenstein, editors, *Findings of the Association for Computational Linguistics: EACL 2023*, pages 265–273, Dubrovnik, Croatia, May. Association for Computational Linguistics.

Svante Wold, Kim Esbensen, and Paul Geladi. 1987. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52.

Pengcheng Yin, Bowen Deng, Edgar Chen, Bogdan Vasilescu, and Graham Neubig. 2018. Learning to mine aligned code and natural language pairs from stack overflow. In *2018 IEEE/ACM 15th international conference on mining software repositories (MSR)*, pages 476–486. IEEE.

Bo Zheng, Li Dong, Shaohan Huang, Wenhui Wang, Zewen Chi, Saksham Singhal, Wanxiang Che, Ting Liu, Xia Song, and Furu Wei. 2021. Consistency regularization for cross-lingual fine-tuning. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3403–3417, Online, August. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023a. Judging llm-as-a-judge with mt-bench and chatbot arena. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

Qinkai Zheng, Xiao Xia, Xu Zou, Yuxiao Dong, Shan Wang, Yufei Xue, Lei Shen, Zihan Wang, Andi Wang, Yang Li, et al. 2023b. Codegeex: A pre-trained model for code generation with multilingual benchmarking on humaneval-x. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5673–5684.

CCL 2024

# 场景图增强的视觉语言常识推理生成

**袁凡, 李丕绩***
计算机科学与技术学院/人工智能学院,
南京航空航天大学
江苏省，南京市，210016
{fanyuan, pjli}@nuaa.edu.cn

## 摘要

视觉语言常识推理是一类旨在理解视觉场景的任务，常用于评估人工智能系统的多模态常识推理能力。然而，可靠的常识推理需要细致的场景理解，而现有的基于预训练模型微调的方法却无法有效地利用具体场景中存在的物体关系信息，因此其推理的合理性存在较大的局限性。为解决上述问题，本研究提出了一种场景图增强的视觉语言常识推理生成框架SGEVL。该框架首先使用图像补丁序列提供视觉信息，并通过一种包含注意力模块的门控机制，赋予大型语言模型理解视觉信息的能力。基于该框架的视觉语言能力，进一步提出了一种无位置信息的场景图生成方法。生成的场景图能够显著提升模型对场景信息的理解，从而引导生成高质量的回答和推理。通过在VCR，VQA-X和e-SNLI-VE数据集上分别实验，实验结果表明本文提出的视觉语言常识推理框架性能优于基线模型。此外，通过消融实验和结果可视化，进一步证明了该框架中每个模块的有效性。

**关键词：** 多模态融合；视觉语言常识推理；场景图生成

# Scene Graph Enhanced Visual Language Commonsense Reasoning Generation

**Fan Yuan, Piji Li***
College of Computer Science and Technology/Artificial Intelligence,
Nanjing University of Aeronautics and Astronautics
Nanjing, Jiangsu 210016, China
{fanyuan, pjli}@nuaa.edu.cn

## Abstract

Visual Language Commonsense Reasoning is a type of task aimed at understanding visual scenes, often used to evaluate the multimodal commonsense reasoning capabilities of AI systems. However, reliable commonsense reasoning needs a detailed understanding of the scene, and existing methods based on fine-tuning pre-trained models are often incapable of effectively utilizing the object relationship information present in specific scenes, thereby limiting the rationality of their reasoning. To address this issue, this study proposes a scene graph enhanced visual language commonsense reasoning generation framework SGEVL. This framework first uses a sequence of image patches to provide visual information and employs a gating mechanism that incorporates an attention mechanism to empower large language models with the ability to understand visual information. Based on the visual language capabilities of this framework, we further propose a method for generating scene graphs without positional

*通讯作者


(c) 2024 中国中文信息学会计算语言学专业委员会

information. The generated scene graphs can significantly enhance the model's understanding of scene information, thereby guiding the generation of high-quality answers and reasoning. Experiments on the VCR, VQA-X, and e-SNLI-VE datasets show that the performance of the proposed framework surpasses that of baseline models. Furthermore, through ablation studies and visualization results, the effectiveness of each module in this framework is further demonstrated.

**Keywords:** Multimodal fusion , Visual language commonsense reasoning , Scene graph generation

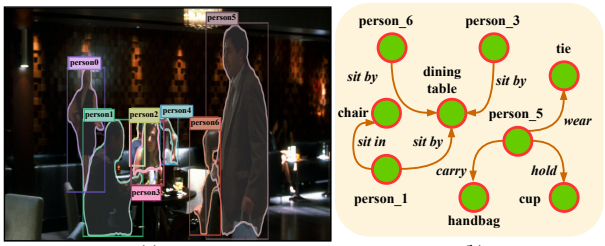
## 1 引言

深度学习模型在视觉-语言任务中已经取得了显著的进步，并展示出其卓越的性能。然而，由于这些模型的“黑箱”特性，其可解释性已经成为了一个日益关注的问题。视觉语言常识推理是一类需要模型从识别和认知两个层面进行推理的视觉推理任务。其中，视觉语言自然语言解释（Vision-language Natural Language Explanation，VL-NLE）(Dua et al., 2021; Kayser et al., 2021; Plüster et al., 2022; Sammani et al., 2022; Whitehouse et al., 2023)近期受到了广泛的关注，其研究中心在于模型是否能够理解视觉场景中的现象和常识，并生成细粒度的推理来解释自身做出的回答或是一些给定的判断。

通过分析，我们发现在VL-NLE多个数据集中的问题、答案和推理解释中包含大量的物体对象及其关系信息。例如，在视觉常识推理数据集（Visual Commonsense Reasoning，VCR）(Zellers et al., 2019)中，有99.7%的问题、97.6%的答案和98.8%的推理解释同时包含了对象和它们的位置关系。平均每一个问题、答案和推理解释的句子都包含超过1个对象和2个以上的关系。特别地，每个推理解释平均包含4个对象和3.8个关系。例如，“He is sitting in a cushioned piece of furniture. He has his briefcase on his lap and he is engrossed in a book.”包含了如下的关系：“He is sitting in furniture”，“briefcase on lap”和“he is engrossed in a book”。因此，对对象及其关系的深入理解在视觉语言常识推理中起着关键的作用。以图1为例，当被问到“what are person1 and person3, and person6 doing?”时，一个没有对象关系提示的模型生成了以下的答案：“person1, person3, and person6 are having a conversation. Because: person1 is smiling and person2 is looking up at person3.”。然而，在实际的场景中，像“dining table”和“cup”这样的关键物体能够明确地帮助我们识别出这是一个用餐的场景，而不仅仅是一个对话的场景。而上述生成的推理过于强调人的因素，忽视了周围环境中的其他对象。因此，考虑到对象关系所生成的推理“they are eating dinner together. Because: they are sitting at a table and have drinks on the table.”则更为精确。

基于以上的观察，我们认为当前的模型在进行视觉语言常识推理时，应当着重于考虑场景中对象的关系。场景图 (Krishna et al., 2017; Qiu et al., 2023; Tang et al., 2020a; Tang et al., 2020b)被认为是一种对对象关系进行建模的有效方法。场景图是由节点和边组成的图结构，其中每个节点代表一个对象，每个有向边表示两端对象之间的关系。例如，如图1 (b)所示，在三元组“⟨person, holding, cup⟩”中，“person”和“cup”是对象，“holding”则是关系。通过提取对象之间的各种位置和语义关系，场景图在某些任务中已经展示出了其潜力 (Yu et al., 2021; Wang et al., 2022b)。考虑到VCR提供的图像无法提供如此详细的位置和关系信息，场景图非常适合在推理过程中作为一座桥梁来填补这个空白。高质量的场景图是提升模型视觉语言常识推理能力的关键。在构建场景图过程中，最常见的方法是首先使用例如Faster-RCNN (Ren et al., 2015)等的目标检测器进行目标检测，然后预测检测到的目标之间的关系。然而，过去的大部分场景图生成方法依赖于对象的位置和边界框信息，采用了复杂的机制和流程 (Dhamo et al., 2020; Dhamo et al., 2021; Johnson et al., 2015)，这对于大多数下游任务来说不是必要的。考虑到上述的不足，我们认为被赋予视觉理解能力的大型语言模型（Large Language Models，LLMs）可以在无显示位置信息的设定下更好地区分两个对象之间的潜在空间关系。

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版




(a) (b)

| | |
|---|---|
| **问题** | what are person1, person3 , and person6 doing? |
| **无场景图辅助的回答和推理** | person1, person3 and person6 are having a conversation.<br>Because: person1 is smiling and person2 is looking up at person3. |
| **有场景图辅助的回答和推理** | they are eating dinner together.<br>Because: they are sitting at a table and have drinks on the table. |
| **真实标签** | they are eating and drinking together.<br>Because: they are all sitting at the same table and they have food and drinks in front of them. |

(c)

Figure 1: 一个视觉语言常识推理的例子。（a）是VCR数据集中的一个场景。（b）是一个从（a）中产生的场景图。在表格（c）中，第一行是问题，第二行是不含场景图辅助的回答和推理，第三行是包含场景图辅助的回答和推理，最后一行是来自数据集的真实标签。

为了解决上述问题，本文提出了一个场景图增强的视觉语言常识推理生成框架SGEVL。为了获取更全面的视觉特征表示，考虑采用CLIP (Radford et al., 2021)模型。CLIP模型的主要结构是一个文本编码器和一个图像编码器，它们分别接收原始文本和图像，然后将输出的图像和文本向量表示映射到一个联合的多模态空间。CLIP模型在训练时采用了对比学习方法，训练其对图像和文本匹配关系的理解能力。对于输入的图像文本对，让正样本对的相似度更高，负样本对的相似度更低。所以CLIP模型拥有较强的视觉图像识别能力以及特征提取能力，能够提供质量较高的视觉特征表示。因此，使用CLIP中的图像补丁序列作为输入，并通过包含注意力模块的门控机制，赋予大型语言模型理解视觉信息的能力。为了实现更可靠的推理感知，我们首先在Visual Genome（VG）数据集 (Krishna et al., 2017)上训练场景图生成，并为视觉语言自然语言解释数据集生成对应的场景图。接着构建prompt模板，结合场景图引导生成高质量的回答和推理。我们在VCR (Zellers et al., 2019)，VQA-X (Park et al., 2016)和e-SNLI-VE (Kayser et al., 2021)三个公共数据集上进行了实验，结果显示我们提出的方法在自动和人工评估指标上都超过了基线方法。消融实验和可视化结果进一步证明了该框架中每个模块的有效性。本文的主要贡献如下：

- 本文结合图像补丁序列和大型语言模型，引入了一种包含注意力模块的门控机制进行模态融合，赋予大型语言模型理解视觉信息的能力。
- 本文提出了一种无位置信息的场景图生成方法和一种基于阈值的场景图筛选方法，并利用筛选后的场景图提供辅助信息，引导生成高质量的回答和推理。
- 本文在三个英文公共数据集上进行了大量的实验，验证了本文提出的框架取得了优于研究进展方法的效果。

## 2 相关工作

场景图包含了详细的对象、属性和关系的语义信息。具体来说，它是由节点和边组成的图结构，其中每个节点代表一个对象，每个有向边表示两端节点之间的关系。场景图的主要表示形式是三元组“⟨主语, 关系, 宾语⟩”，其中主语和宾语用节点表示，而关系则用有向边表示。场景图生成（Scene Graph Generation，SGG） (Lu et al., 2016; Li et al., 2017a; Zhang et al., 2017)是一项从图像场景中提取此类语义表示的任务。现有的场景图生成方法可以分为两类。第一类包括两个阶段：首先使用Faster-RCNN (Ren et al., 2015)等目标检测方法进行目标检测，然后在此基础上进行关系预测 (Lu et al., 2016; Dai et al., 2017; Liao et al., 2019)。考虑到图像中的元素可能表示其他元素的上下文，第二种方法采用对象和对象关系的联合预测 (Li et al., 2017a; Li et al., 2017b; Xu et al., 2017)。一些工作专注于场景图中的长尾问题 (Zhang et al.,


（c）2024 中国中文信息学会计算语言学专业委员会

2017; Krishna et al., 2019; Zareian et al., 2020)，并提出了无偏场景图生成 (Yan et al., 2020; Wang et al., 2020)。此外，一些研究已经提出了无位置场景图生成的概念 (Özsoy et al., 2023)，其目标是在不使用边界框等位置信息的情况下进行场景图生成。尽管现有的研究已经能够生成相对高质量的场景图，但要实现其实用性仍面临着巨大的挑战。

视觉语言自然语言解释（Vision-language Natural Language Explanation，VL-NLE）(Dua et al., 2021; Kayser et al., 2021; Plüster et al., 2022; Sammani et al., 2022; Whitehouse et al., 2023)是自然语言解释任务（Natural Language Explanation，NLE）在多模态领域的扩展。它旨在通过生成对人类友好且细粒度的自然语言句子来解释黑盒模型的决策过程。e-ViL (Kayser et al., 2021)是一个针对VL-NLE任务的评估基准，其中包含VQA-X (Park et al., 2018)、VCR (Zellers et al., 2019)和e-SNLI-VE (Kayser et al., 2021)三个数据集。视觉常识推理（Visual Commonsense Reasoning，VCR）(Zellers et al., 2019)是其中较为复杂的一项多模态任务。VCR数据集提供了大量的图像和相关问题，并基于这些信息推断出答案和解释。每条数据都包含四个备选答案和四个备选推理选项。与经典的视觉问题回答任务不同，VCR提供的选项通常涉及到多组对象关系或事件的全面描述，要求模型对场景信息和常识信息有深刻的理解，因此同时做出正确的选择是具有挑战性的。为了评测模型的生成推理能力，e-ViL将VCR任务从分类任务转化为生成任务。过去已经有许多工作 (Dua et al., 2021; Kayser et al., 2021; Whitehouse et al., 2023)致力于提高模型的解释能力。例如，e-UG (Kayser et al., 2021)使用UNITER (Chen et al., 2020)作为预测模块进行解释预测，NLX-GPT (Sammani et al., 2022)通过训练一个蒸馏的GPT-2 (Radford et al., 2019)来结合答案和解释的生成。这些方法在捕捉场景中对象之间精确关系以进行推理的能力上仍存在困难，而解决这个问题正是本文的主要目标。

## 3 方法

为了增进模型生成合理回答和解释的能力，本文提出了一种场景图强化的视觉语言常识推理生成方法SGEVL。该方法基于输入的图像和问题，首先根据图像生成对应的场景图，然后利用场景图来改善模型视觉语言常识推理的性能。整体框架如图2所示。其中，定义$I$为给定的图像，$Q$为相应的问题，$O = (o_1, o_2, \ldots, o_n)$为对应图像中的对象，其中$n$表示对象的数量。在无位置的场景图构建阶段，将图像$I$的补丁序列$P = (p_1, p_2, \ldots, p_m)$作为视觉输入$X_{\mathrm{v}}$，将包含对象的文本提示prompt作为文本输入$X_{\mathrm{t}}$。目标是预测一个场景图$G = (V, E)$，其中$V$和$E$分别代表实体节点和关系。随后，在生成视觉语言常识推理的答案和解释时，使用图像$I$的补丁序列$P = (p_1, p_2, \ldots, p_m)$作为视觉输入$X_{\mathrm{v}}$，对问题$Q$和场景图$G$构建文本提示prompt作为输入$X_{\mathrm{t}}$。目标是生成答案和解释$(I, G, Q) \rightarrow (A, R)$，其中$A$代表答案，$R$表示解释。我们期望生成的$A, R$在$G$的帮助下是合理并符合常识的。

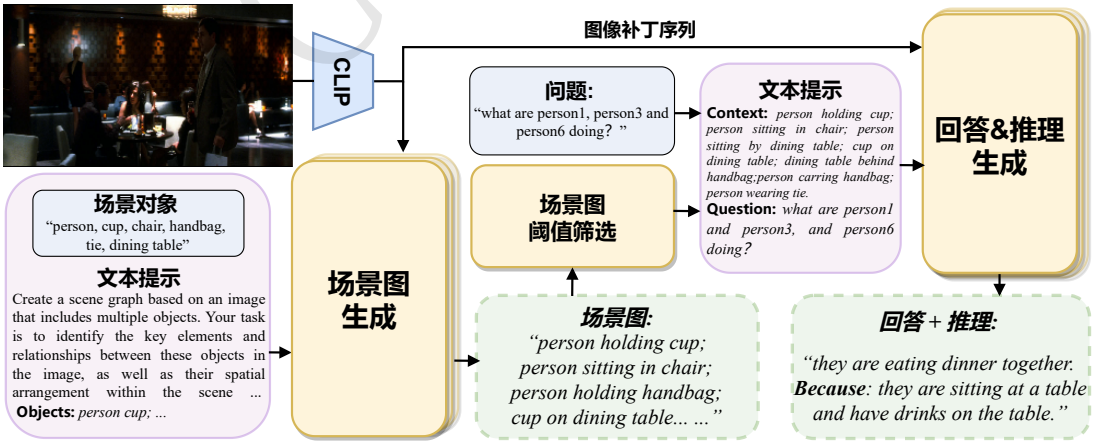



Figure 2: SGEVL框架的结构图。该框架首先根据图像的补丁序列和物体提示生成场景图。然后，结合问题和图像，生成符合常识的回答和解释。


(c) 2024 中国中文信息学会计算语言学专业委员会

### 3.1 视觉语言模态融合

由于大型语言模型是文本单模态模型，无法利用图像场景信息，因此本文采用了一种基于单头注意力的门控机制对为大型语言模型提供多模态能力。为了使不同模态的特征进行有效的融合，并更大程度的保留视觉信息，我们使用CLIP模型 (Radford et al., 2021)的图像补丁序列输出$P = (p_1, p_2, \ldots, p_m)$作为视觉输入$X_\mathrm{v}$，并使用一个线性层将其投影到与文本输入相同的隐层维度，定义为视觉隐层向量$\mathbf{H}_\mathrm{v}$。接着，使用一个单头跨模态注意力模块对视觉输入和文本输入进行模态融合获得跨模态隐层向量$\mathbf{H}_\mathrm{c}$:

$$\mathbf{H}_\mathrm{t} = \mathbf{TextEncoder}(X_\mathrm{t}), \tag{1}$$

$$\mathbf{H}_\mathrm{v} = \mathbf{W_h} \cdot X_\mathrm{v}, \quad \text{s.t. } X_\mathrm{v} = \mathbf{CLIP}(I), \tag{2}$$

$$\mathbf{H}_\mathrm{c} = \mathbf{Attention}\left(\mathbf{H}_\mathrm{t}\mathbf{W^Q}, \mathbf{H}_\mathrm{v}\mathbf{W^K}, \mathbf{H}_\mathrm{v}\mathbf{W^V}\right), \tag{3}$$

其中，文本隐层向量$\mathbf{H}_\mathrm{t}$作为查询向量，视觉隐层向量$\mathbf{H}_\mathrm{v}$作为键值向量和值向量，$\mathbf{W^Q}$，$\mathbf{W^K}$，$\mathbf{W^V}$和$\mathbf{W_h}$分别表示可学习的参数。

为了更好地利用大型语言模型的文本语义理解能力，我们引入了一种门控机制 (Zhang et al., 2023)帮助平衡两个不同的模态。首先，对跨模态隐层向量$\mathbf{H}_\mathrm{c}$和文本隐层向量$\mathbf{H}_\mathrm{t}$在特征维度上进行拼接。接着，使用线性层将其维度调整至与隐层向量相同，并通过sigmoid函数计算出门控参数$\lambda$。最后，使用$\lambda$对跨模态隐层向量$\mathbf{H}_\mathrm{c}$和文本隐层向量$\mathbf{H}_\mathrm{t}$加权相加获得融合隐层向量$\mathbf{H}_\mathrm{f}$。过程可以表示为:

$$\lambda = \mathbf{Sigmoid}\left(\mathbf{W_t} \cdot \mathbf{H}_\mathrm{t} \oplus \mathbf{W_v} \cdot \mathbf{H}_\mathrm{c}\right), \tag{4}$$

$$\mathbf{H}_\mathrm{f} = (1 - \lambda) \cdot \mathbf{H}_\mathrm{t} + \lambda \cdot \mathbf{H}_\mathrm{c}, \tag{5}$$

$\mathbf{W_t}$和$\mathbf{W_v}$分别表示可学习的参数。融合隐层向量$\mathbf{H}_\mathrm{f}$作为模态联合输入参与模型的训练与推理。

### 3.2 无位置信息的场景图生成

无位置信息的场景图生成使用Visual Genome（VG）数据集 (Krishna et al., 2017)进行训练。在VG数据集中，每一张图像包含不同数量的场景图，而这些场景图中的对象也在大小上有所不同。在一般的认知中，特别小的对象，如“hair”和“finger”，通常会包含更普遍的关系信息，在其关联的场景图中能提供的信息较少。而较大的对象往往能够引领场景中的更多关系信息，在推断中起着更关键的作用。因此，为了提升模型训练效率，我们对VG数据集进行了一定程度的数据清理。根据不同场景图三元组中主语对象和宾语对象的平均大小对每个图像对应的所有场景图进行排序，并选择前50个进行训练。这些场景图已经可以囊括大部分有效的对象，并有效排除部分信息量较小的三元组。

由于大型语言模型无法通过边界框等显示位置数据定位对象，因此，为了训练模型获得无位置信息的场景图生成能力，我们不使用场景中各个对象的边界框等显示位置数据作为输入。但是，为了防止图像场景中存在相同命名的对象影响模型的预测效果，根据数据集中所选对象从左往右出现的位置给定一个序号。例如，场景中存在多个“person”对象，将左边的对象定义为“person0”，右边的对象定义为“person1”，并以此类推。同时，我们也加入了个别无具体关系的对象，构成“⟨主语, None, 宾语⟩”的三元组，以训练模型应对无关对象的情况。

具体来说，使用基于Transformer (Vaswani et al., 2017)的大型语言模型FLAN-T5 (Chung et al., 2022)作为主干模型，它是一个编码器-解码器架构。受到Vision Transformer (Dosovitskiy et al., 2021; Liu et al., 2021)在视觉领域成功的启发，我们使用经过训练的CLIP模型提取图像的补丁序列作为视觉输入。将标签中的场景图拼接作为待预测的场景图三元组序列$G = \{sub_1\ obj_1\ rel_1;\ \ldots\ ; sub_\sigma\ obj_\sigma\ rel_\sigma\}$，其中$\sigma$表示三元组的数量。并基于此将筛选过的场景图中的对象按照$X_\mathrm{o} = \{sub_1\ obj_1; sub_2\ obj_2;\ \ldots\ ; sub_\sigma\ obj_\sigma\}$的形式构成文本输入。为了能够利用大型语言模型理解指令的能力，我们构建了文本提示“Create a scene graph based on an image that includes multiple objects. The task is to identify the key elements and relationships between these objects in the image, as well as their spatial arrangement within the scene. Objects:$\{X_\mathrm{o}\}$ Scene:”。然后将文本提示作为文本输入$X_\mathrm{t}$，图像补丁序列作为视觉输入$X_\mathrm{v}$，经过第3.1节的模态融合模块后生成对应的场景图$G$。


(c) 2024 中国中文信息学会计算语言学专业委员会

由于预测场景图的方式是从序列到序列的，因此，采用交叉熵损失作为无位置信息的场景图生成的训练损失：

$$\mathcal{L}_{SG} = -\sum_{l}^{L} \log P\left(g_l \mid g_{<l}, X_{\mathrm{t}}; X_{\mathrm{v}}\right), \tag{6}$$

其中，$g_l$是第$l$个生成的词，$X_{\mathrm{t}}$包含对象信息的$X_{\mathrm{o}}$文本提示，$X_{\mathrm{v}}$表示视觉输入。

### 3.3 视觉语言常识推理生成

在视觉语言常识推理阶段，视觉语言自然语言解释（VL-NLE）的目标是先生成对应的回答，再生成能够帮助解释回答的推理。为了提升模型推理的一致性，我们将这两个离散的任务合并为一个，即同时生成回答和推理。具体来说，同样使用基于Transformer (Vaswani et al., 2017)的大型语言模型FLAN-T5 (Chung et al., 2022)作为主干模型。使用文本提示："Context:{$G$} Question:{$Q$}"作为文本输入$X_{\mathrm{t}}$。其中，$G = \{sub_1\ obj_1\ rel_1;\ \ldots\ ; sub_\sigma\ obj_\sigma\ rel_\sigma\}$表示在第3.2节中生成的场景图三元组序列。问题$Q$表示为$Q = \{q_1, q_2, \ldots, q_h\}$，$h$是输入问题的长度。对于视觉输入$X_{\mathrm{v}}$，我们使用与构建场景图时相同的CLIP图像补丁序列，并使用第3.1节中的跨模态注意力和门控机制进行模态融合。

需要注意的是，在场景图生成阶段之后，每个图像都会得到一个相应的由一系列的三元组组成的场景图。然而，并非所有的三元组都是准确的，不准确的三元组不仅对生成过程没有正面的帮助，而且可能还引入额外的噪音，对推理造成误导。因此，为了确保生成更准确和合理的推理，筛选出高质量生成的三元组至关重要。本文采用了一种基于阈值的选择方法：对于原始图像$I$和生成的场景图$G$，我们使用经过训练的CLIP模型计算每个三元组和图像区域的标准化相似度得分。输出的得分按降序排序，得分越高表示相似度越高，表示它们更符合图像场景。我们为选择过程设定一个阈值，然后从有序集合中持续选择具有最高置信度得分的三元组。即当累计的置信度得分之和小于阈值时，将不断继续选择新的三元组，直到达到阈值为止。

同时，我们将真实标签中的回答和推理以"{$A$} Because:{$R$}"的形式组合作为$(Q \rightarrow AR)$生成的训练标签。考虑到VL-NLE的原始设定还包含了两个子任务：$(Q \rightarrow A)$和$(QA \rightarrow R)$，我们也为此构造了相应的文本提示模板。更具体地说，对于子任务$(Q \rightarrow A)$（即输入问题，生成回答），设置"Context:{$G$} Question:{$Q$} Answer:"的输入模板和"Answer:{$A$}"的真值标签格式。对于子任务$(QA \rightarrow R)$（即输入问题和回答，生成推理），设置了"Context:{$G$} Question:{$Q$} Answer:{$A$} Because:"的输入模板和"Explanation:{$R$}"的真值标签格式。其中，$A$和$R$分别表示生成的回答和推理解释。最后，训练中的生成损失被定义为以下的交叉熵损失：

$$\mathcal{L}_{E} = -\sum_{i}^{I} \log P\left(y_i \mid y_{<i}, X_{\mathrm{t}}; X_{\mathrm{v}}\right), \tag{7}$$

其中，$y_i$是生成的词，$X_{\mathrm{t}}$是包含问题$Q$和场景图$G$的文本输入，$X_{\mathrm{v}}$表示视觉输入。

## 4 实验设置

### 4.1 实验对比模型

为了评估我们提出的模型框架，我们与以下模型进行比较：

- **e-UG** (Kayser et al., 2021)：e-UG将GPT-2与UNITER结合，使用Faster R-CNN获取区域的视觉特征，并将位置特征编码为视觉输入。通过将图像区域和问题词的嵌入预置到文本问题和预测答案中，并利用UNITER的上下文嵌入输出来微调GPT-2。

- **OFA-X** (Plüster et al., 2022)：OFA-X将OFA (Wang et al., 2022a)模型作为主干模型，它是一种标准的编码器-解码器Transformer架构。给模型提供问题和四个不同的选择，根据模型对每一个选项输出的的可能性选择正确的答案和解释。

- **NLX-GPT** (Sammani et al., 2022)：NLX-GPT是一种编码器-解码器模型，由对图像进行编码的视觉主干模型和精简的GPT-2组成。首先在图像标注任务上预训练精简的GPT-2，然后在VL-NLE数据集上进行微调。


(c) 2024 中国中文信息学会计算语言学专业委员会

- **UMAE** (Whitehouse et al., 2023)：UMAE同样基于OFA模型。为多个数据集设计了相应的文本提示，并在VL-NLE任务上进行多任务训练。

### 4.2 评测指标

为了评估生成的场景图的质量，并衡量它们对视觉语言常识推理生成的帮助程度，我们分别对场景图生成和视觉语言常识推理生成在相应指标下进行评估 (Özsoy et al., 2023; Yu et al., 2021)。

对于无位置信息场景图生成，我们采用启发式树搜索（Heuristic Tree Search，HTS）(Özsoy et al., 2023)来评估生成质量，该方法用于计算图之间的重叠程度并获取召回率分数。在预测阶段，生成的场景图以序列形式呈现。我们首先将预测的序列和标签序列转换为图结构，然后使用这种基于树的搜索算法来计算召回率R@50和100。

对于视觉语言常识推理生成，我们采用了不同的评估方法，包括自动评估和人工评估。在自动评估中，我们使用了自然语言生成的评测指标：BLEU (Papineni et al., 2002)，METEOR (Banerjee and Lavie, 2005)，ROUGE-L (Lin, 2004)，CIDEr (Vedantam et al., 2015)，SPICE (Anderson et al., 2016)和BERTScore F1 (Zhang et al., 2020)。同时，还使用了e-ViL指标 (Kayser et al., 2021)进行评测。e-ViL指标由$S_T$，$S_E$和$S_O$组成，分别代表任务准确度，n-gram综合分数，以及总分（$S_T$和$S_E$的乘积）。$S_E$是通过计算ROUGE-L，METEOR，CIDEr，SPICE和额外的BERTScore F1的调和平均数得到的。为了与之前的工作 (Yu et al., 2021; Whitehouse et al., 2023; Sammani et al., 2022; Plüster et al., 2022))在相同的尺度下进行比较，我们同样使用过滤后的生成结果来评估我们模型的性能。即在过滤过程中，采用了0.92的BERTscore阈值来判断回答部分的正确性：BERTscore低于0.92的回答被认为是错误的，而达到0.92的回答被认为是正确的。因此，经过过滤后的回答被认为是正确的，并进一步评估其推理解释。此外，我们还使用了人工评估，方法与之前的工作 (Kayser et al., 2021)相同：随机从测试集中选择300个答案正确的例子，注释者需要从“yes”，“weak yes”，“weak no”和“no”中选择一个作为推理是否能够解释回答的评测结果。

### 4.3 实验参数设置

在实验中，我们使用训练过的CLIP模型[0]的补丁序列输出作为图像输入，用预训练过的FLAN-T5-large[1]的参数来初始化模型。在场景图生成阶段，我们使用大小为8的批次大小训练模型，共迭代5个轮次。在视觉语言常识推理生成阶段，我们使用大小为16的批次大小训练模型，共迭代4个轮次。在两个阶段中，都使用AdamW优化器 (Loshchilov and Hutter, 2019)来训练模型，学习率都被设定为$5e^{-5}$，热身步数分别设置为1k和2k步。训练中的模型每经过5k步进行一次验证，并保存最好的模型进行评测。实验均在8张NVIDIA RTX 3090 24Gb显卡上进行。

| 模型 | 边界框 | R@50 | R@100 |
|---|---|---|---|
| SGTR | ✓ | 30.38 | 34.85 |
| Pix2SG | ✗ | 24.81 | 26.66 |
| **ours** | ✗ | 25.93 | 43.61 |

Table 1: Visual Genome数据集上的场景图生成在R@k指标上的实验结果。

## 5 结果与分析

### 5.1 场景图生成的实验结果

与大多数场景图生成模型不同，本文提出的方法采用了无位置设定，即不提供边界框等显示的位置信息。因此，我们仅采用启发式树搜索方法对场景图进行评测，结果如表1所示。SGTR (Li et al., 2022)是一种传统方法下的端到端的场景图生成模型，它分别预测实体节点和谓词节点，通过连接两种节点，生成二分图进而构建场景图。Pix2SG (Özsoy et al., 2023)是最早也是目前唯一在无位置设定下提出场景图生成的模型。它基于Pix2Seq骨干网

[0]https://github.com/jianjieluo/OpenAI-CLIP-Feature
[1]https://huggingface.co/google/flan-t5-large


(c) 2024 中国中文信息学会计算语言学专业委员会

络 (Chen et al., 2022)，这是一种能够将物体检测任务表示为序列预测的多模态模型。实验结果表明，SGEVL与传统的场景图生成模型相比，能够在R@100上得到显著的提升。而在无位置信息的设定下，本文构建的场景图生成模型在R@50和R100上都能够超过Pix2SG模型。这表明赋予多模态能力后的大型语言模型的视觉语义理解能力得到了提升，并能够有效地识别场景中的对象以及其关联信息。

### 5.2 视觉语言常识推理的实验结果

我们分别在VCR，VQA-X和e-SNLI-VE三个公开数据集上进行了实验，并与基准模型进行了比较，结果如表2所示。下面将根据不同数据集分别进行结果分析。

#### 5.2.1 VCR数据集的实验结果

根据表2所示，现有的一些工作（例如e-UG，OFA-X和NLX-GPT）在VCR数据集上表现出了相近的性能。他们在BLEU和SPICE这样的n-gram指标上得分较低，表明这些模型生成的句子结构与标签相比存在一定差距。与这些模型相比，UMAE的性能取得了更加显著提升，包括将BLEU-4从6.6%提高到13.4%，展现出了更强的句子结构一致性。然而，与其他模型相比，它在$S_T$指标上的得分下降了大约10%，表明其生成结果的准确性存在欠缺。

本文提出的方法与先前的工作相比取得了显著的提升。具体来说，加入了场景图强化后，模型在METEOR、ROUGE-L、CIDEr和BERTScore上分别达到了20.9、34.9、57.5和90.1的得分，超过了基准模型的表现，这意味着场景图提供的对象和对象间的关系有助于模型生成与标签语义更相近的解释。同时，在e-ViL指标上，SGEVL在$S_O$和$S_E$上分别取得了28.3和45.6的分数，表明生成的推理解释能够达到更高的标准。

#### 5.2.2 VQA-X数据集的实验结果

在VQA-X数据集上的实验结果如表2所示。SGEVL在n-gram评测指标和BERTScore指标上取得了相对较高的分数。其中，METEOR和SPICE得到了最高的分数，分别为24.0和24.3，并在$S_E$分数上超过基准模型，得到了50.6。但是，由于VQA-X数据集中提供的回答和推理的句子长度较短（通常以单词或短语的方式组成），而我们的模型倾向于生成长度更长，更加完整的推理过程，这会间接导致评测分数的下降。因此在结果中本文的方法在个别指标上略低于基线方法。但从综合评估的角度，SGEVL仍然表现出了良好的效果。

#### 5.2.3 e-SNLI-VE数据集的实验结果

e-SNLI-VE数据集上的实验结果也同样表明了我们提出的框架的有效性。根据表2中的结果，本文的提出的方法在所有指标上都取得了较好的分数，特别是在METEOR和ROUGE-L等指标上。此外，与基线工作相比，我们的方法在综合得分$S_O$上实现了7%的提升。这进一步证实了我们提出的框架能够把握住视觉场景中的关系信息，生成符合场景信息的解释。

#### 5.2.4 人工评测

为了更好地评估视觉语言常识推理的生成质量，本文从测试样集中随机抽取了部分经过BERTScore过滤和未经过过滤的测试样例进行了人工评测。评测人员需要根据生成的推理能否佐证回答，从“yes”，“weak yes”，“weak no”和“no”四个选项中对每一个样例进行评价。评测结果如表3所示。在经过过滤的样例中，有63.1%样例的推理部分被认为很好地证明和解释了回答部分，只有5.4%的样例在推理和解释上没有得到人工评测的认可。这表明从人类认知的角度，绝大部分的样例都是符合常识和具有解释性的。而在未经过过滤的样例中，“weak no”和“weak yes”的比例出现了增长，而被选为“no”的比例依然较小，这也表明在回答不完全正确的情况下，推理的质量依然能够保持一定的稳定性。

### 5.3 消融实验

#### 5.3.1 融合不同的视觉特征

我们进行了消融实验以了解使用不同视觉特征的效果。结果如表4所示。其中，DETR (Carion et al., 2020)是一个用于目标检测的端到端的基于Transformer的模型。CLIP用于分别提取补丁序列特征和全局特征。可以观察到，相比于补丁序列特征，DETR在这些指标上只显示出轻微的落后，而全局特征在召回率上则表现出明显的不


(c) 2024 中国中文信息学会计算语言学专业委员会

| | e-ViL Scores | | | *n*-gram Scores | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ***VCR*** | $S_O$ | $S_T$ | $S_E$ | B1 | B2 | B3 | B4 | M | R-L | C | S | BS |
| e-UG | 19.3 | 69.8 | 27.6 | 20.7 | 11.6 | 6.9 | 4.3 | 11.8 | 22.5 | 32.7 | 12.6 | 79.0 |
| OFA-$X_{VCR}$ | 23.0 | **71.2** | 32.4 | 24.5 | 14.4 | 9.1 | 6.1 | 12.2 | 25.1 | 48.5 | 18.8 | 79.8 |
| OFA-$X_{MT}$ | 19.2 | 62.0 | 30.9 | 22.3 | 13.0 | 8.0 | 5.2 | 11.3 | 24.3 | 44.6 | 17.8 | 79.3 |
| NLX-GPT | - | - | 32.6 | 24.7 | 15.0 | 9.6 | 6.6 | 12.2 | 26.4 | 46.9 | 18.8 | 80.3 |
| $UMAE_{VCR}$ | 22.5 | 56.6 | 39.8 | - | - | - | 12.3 | 16.7 | 28.9 | 48.2 | **27.4** | 81.8 |
| $UMAE_{MT}$ | 22.8 | 56.6 | 40.2 | 31.4 | 22.9 | 17.6 | **13.4** | 17.5 | 29.5 | 47.3 | 26.5 | 81.9 |
| **SGEVL(ours)** | **26.1** | 61.9 | **45.3** | **35.7** | **24.2** | **17.3** | 12.9 | **20.7** | **34.5** | **55.6** | **27.4** | **89.4** |
| ***VQA-X*** | | | | | | | | | | | | |
| e-UG | 36.4 | 80.5 | 45.3 | 57.3 | 42.7 | 31.4 | 23.2 | 22.1 | 45.7 | 74.1 | 20.1 | 87.0 |
| OFA-$X_{MT}$ | **45.5** | **92.6** | 49.2 | 64.0 | 49.4 | **37.6** | **28.6** | 23.1 | 51.0 | 110.2 | 22.6 | 86.8 |
| NLX-GPT | 40.6 | 83.0 | 49.0 | **64.2** | **49.5** | **37.6** | 28.5 | 23.1 | **51.5** | **110.6** | 22.1 | **86.9** |
| $UMAE_{MT}$ | 31.5 | 77.6 | 40.6 | 47.5 | 31.4 | 21.4 | 14.6 | 20.2 | 35.1 | 50.3 | 19.1 | 85.4 |
| **SGEVL(ours)** | 43.5 | 87.5 | **49.8** | **64.2** | 48.6 | 36.4 | 27.0 | **23.9** | 44.8 | 103.8 | **23.1** | 86.7 |
| ***e-SNLI-VE*** | | | | | | | | | | | | |
| e-UG | 36.0 | **79.5** | 45.3 | 30.1 | 19.9 | 13.7 | 9.6 | 19.6 | 27.8 | 85.9 | 34.5 | 81.7 |
| OFA-$X_{MT}$ | 35.6 | 78.9 | 45.1 | 32.4 | 21.8 | 15.2 | 10.8 | 17.9 | 31.4 | 108.2 | 32.8 | 80.4 |
| NLX-GPT | 34.6 | 73.9 | 46.9 | 37.0 | 25.3 | 17.9 | 12.7 | 18.8 | 34.2 | **117.4** | 33.6 | 80.8 |
| **SGEVL(ours)** | **41.7** | 73.9 | **56.5** | **39.8** | **27.5** | **17.9** | **12.8** | **24.9** | **39.7** | 114.7 | **43.9** | **89.3** |

Table 2: VCR，VQA-X和e-SNLI-VE数据集上的过滤评测结果。B，M，R-L，C，S和BS分别表示BLEU，METEOR，ROUGE-L，CIDEr，SPICE和BERTScore F1。模型下标的**VCR**表示该模型仅在VCR数据集上训练，而**MT**表示该模型经过多任务训练。

| | No | Weak No | Weak Yes | Yes |
|---|---|---|---|---|
| 过滤后 | 5.4% | 5.2% | 26.3% | 63.1% |
| 过滤前 | 9.5% | 29.6% | 29.1% | 31.8% |

Table 3: 人工评测结果。“过滤后”表示仅对回答正确的样例进行推理生成的评测。“过滤前”表示同时对所有样例的回答和推理的评测。

足。我们认为这主要是因为在获取全局特征的过程中，由于维度空间的的压缩导致了信息的丢失，而补丁序列特征可以为大型语言模型提供更原始，更完整的视觉信息。

为了验证我们的模态融合机制的有效性，我们还对是否使用模态融合机制进行了进一步的分析，并通过表5补充了我们的研究。其中，“$w/o\ I$”表示仅加入了场景图，未加入图像和模态融合机制。实验结果显示，视觉模态的融合显著地提升了结果，进一步证明了其有效性。

### 5.3.2 选择不同的场景图筛选阈值

我们还对不同的场景图选择方法进行了消融实验。结果显示在表6中。可以观察到，当阈值设定为0.8时，过滤后的评测结果可以达到最优的性能。然而，当阈值降低或提高时，性能显示出下降趋势，甚至低于不使用场景图的水平。不仅如此，提供较少的场景图将会导致场景信息的不足，而提供更多质量较差的场景图将导致噪声的增加，这都将降低最终推理生成的表

| 视觉特征 | 维度 | R@20 | R@50 | R@100 |
|---|---|---|---|---|
| DETR | (100, 256) | 10.07 | **26.00** | 43.39 |
| $CLIP_{global}$ | (1, 512) | 6.39 | 15.82 | 27.56 |
| $CLIP_{patch}$ | (49, 2048) | **10.92** | 25.93 | **43.61** |

Table 4: 使用不同视觉特征输入在Visual Genome数据集上进行场景图生成的消融实验。


(c) 2024 中国中文信息学会计算语言学专业委员会

| | $S_O$ | $S_T$ | $S_E$ | B1 | B4 | R-L | BS |
|---|---|---|---|---|---|---|---|
| SGEVL ($w/o\ I$) | 13.7 | 43.3 | 31.7 | 27.9 | 5.6 | 25.0 | 77.8 |
| SGEVL | 26.1 | 61.9 | 45.3 | 35.7 | 12.9 | 34.5 | 89.4 |

Table 5: 模态融合的消融实验结果。

| 方法 | B1 | B2 | B3 | B4 | M | R-L | C | S | BS |
|---|---|---|---|---|---|---|---|---|---|
| SGEVL ($w/o$ 场景图) | 35.3 | 23.8 | 16.9 | 12.5 | **20.7** | **34.5** | 53.7 | 26.8 | **90.0** |
| SGEVL ($w/$ 0.7) | 35.4 | 23.7 | 16.8 | 12.3 | 20.0 | 33.6 | 49.8 | 26.5 | 89.8 |
| SGEVL ($w/$ 0.8) | **35.7** | **24.2** | **17.3** | **12.9** | **20.7** | **34.5** | **55.6** | **27.4** | **90.0** |
| SGEVL ($w/$ 0.9) | 35.3 | 23.8 | 16.9 | 12.4 | 20.0 | 33.8 | 50.6 | 26.5 | 88.5 |
| SGEVL ($w/$ 1.0) | 34.9 | 22.9 | 16.2 | 11.7 | 19.8 | 33.8 | 49.9 | 26.4 | 88.6 |

Table 6: 使用不同场景图筛选阈值的消融实验。

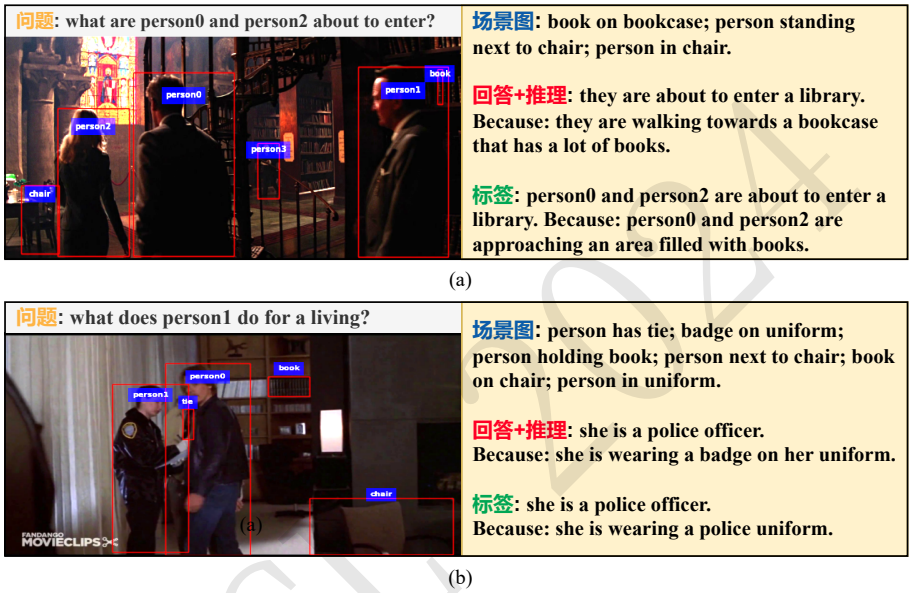



Figure 3: SGEVL在VCR数据集上的案例分析。

现。生成的场景图的质量对推理的质量存在相关性，筛选更准确的场景图有助于提升推理的合理性。因此，根据结果本文选用0.8作为最终的筛选阈值。

### 5.4 可视化分析

我们对部分样例进行可视化并为进行了分析，以对所提出的方法有一个更加清晰的认知。我们在VCR数据集上的可视化结果如图3所示。可以看到，生成的场景图三元组为核心对象之间的详细关系提供了信息。在图3(a)的例子中，生成的“book on bookcase”增强了这是一个图书馆的场景的可能性，推理模型正确地应用了这个三元组生成了“they are about to enter a library. Because: they are walking towards a bookcase that has a lot of books”。生成的推理能够较好地反映场景中的情况，能够提供关键的信息，并对回答生成符合常理的解释。图3(b)中展示的例子中，推理模型通过生成的“person has tie”和“person in uniform”三元组，帮助推理模型生成了一个与实际情况密切匹配的推理解释。

## 6 结论

本文主要研究了如何增强模型在视觉语言常识推理能力的问题。为解决在常识推理过程中大型语言模型缺乏对场景关系的感知的问题，本文提出了一个场景图增强的的推理生成框架。这个解耦的框架将无位置信息场景图生成和视觉语言常识推理生成分离为两个阶段。在多个数

据集上的实验显示出本文的方法在与基线方法的比较中取得了更好的效果，证明了本文的方法所生成的场景图有助于产生更合理的推理，从而解释和验证回答的准确性。不仅如此，本文还针对不同模块进行了消融实验，以验证框架中各个部分的有效性。在未来的工作中，我们将探索更全面的视觉关系信息挖掘方法，并加强模型的逻辑推理能力。

## 7 致谢

该研究得到了国家自然科学基金（62106105）、CCF-百度松果基金(CCF-Baidu202307)、CCF-智谱大模型基金(CCF-Zhipu202315)、南京航空航天大学科研启动基金（YQR21022）和南京航空航天大学高性能计算平台的支持。

## 参考文献

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: semantic propositional image caption evaluation. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V*, volume 9909 of *Lecture Notes in Computer Science*, pages 382–398. Springer.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare R. Voss, editors, *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, pages 65–72. Association for Computational Linguistics.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, volume 12346 of *Lecture Notes in Computer Science*, pages 213–229. Springer.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: universal image-text representation learning. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, volume 12375 of *Lecture Notes in Computer Science*, pages 104–120. Springer.

Ting Chen, Saurabh Saxena, Lala Li, David J. Fleet, and Geoffrey E. Hinton. 2022. Pix2seq: A language modeling framework for object detection. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

Bo Dai, Yuqi Zhang, and Dahua Lin. 2017. Detecting visual relationships with deep relational networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3298–3308. IEEE Computer Society.

Helisa Dhamo, Azade Farshad, Iro Laina, Nassir Navab, Gregory D. Hager, Federico Tombari, and Christian Rupprecht. 2020. Semantic image manipulation using scene graphs. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 5212–5221. Computer Vision Foundation / IEEE.

Helisa Dhamo, Fabian Manhardt, Nassir Navab, and Federico Tombari. 2021. Graph-to-3d: End-to-end generation and manipulation of 3d scenes using scene graphs. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 16332–16341. IEEE.


(c) 2024 中国中文信息学会计算语言学专业委员会

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Radhika Dua, Sai Srinivas Kancheti, and Vineeth N. Balasubramanian. 2021. Beyond VQA: generating multi-word answers and rationales to visual questions. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2021, virtual, June 19-25, 2021*, pages 1623–1632. Computer Vision Foundation / IEEE.

Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2015. Image retrieval using scene graphs. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3668–3678. IEEE Computer Society.

Maxime Kayser, Oana-Maria Camburu, Leonard Salewski, Cornelius Emde, Virginie Do, Zeynep Akata, and Thomas Lukasiewicz. 2021. e-vil: A dataset and benchmark for natural language explanations in vision-language tasks. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 1224–1234. IEEE.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1):32–73.

Ranjay Krishna, Vincent S. Chen, Paroma Varma, Michael S. Bernstein, Christopher Ré, and Li Fei-Fei. 2019. Scene graph prediction with limited labels. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 2580–2590. IEEE.

Yikang Li, Wanli Ouyang, Xiaogang Wang, and Xiaoou Tang. 2017a. Vip-cnn: Visual phrase guided convolutional neural network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 7244–7253. IEEE Computer Society.

Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. 2017b. Scene graph generation from objects, phrases and region captions. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 1270–1279. IEEE Computer Society.

Rongjie Li, Songyang Zhang, and Xuming He. 2022. SGTR: end-to-end scene graph generation with transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 19464–19474. IEEE.

Wentong Liao, Bodo Rosenhahn, Ling Shuai, and Michael Ying Yang. 2019. Natural language guided visual relationship detection. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 444–453. Computer Vision Foundation / IEEE.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9992–10002. IEEE.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Cewu Lu, Ranjay Krishna, Michael S. Bernstein, and Li Fei-Fei. 2016. Visual relationship detection with language priors. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*, volume 9905 of *Lecture Notes in Computer Science*, pages 852–869. Springer.

Ege Özsoy, Felix Holm, Tobias Czempiel, Nassir Navab, and Benjamin Busam. 2023. Location-free scene graph generation. *CoRR*, abs/2303.10944.


(c) 2024 中国中文信息学会计算语言学专业委员会

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2016. Attentive explanations: Justifying decisions and pointing to the evidence. *CoRR*, abs/1612.04757.

Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2018. Multimodal explanations: Justifying decisions and pointing to the evidence. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 8779–8788. Computer Vision Foundation / IEEE Computer Society.

Björn Plüster, Jakob Ambsdorf, Lukas Braach, Jae Hee Lee, and Stefan Wermter. 2022. Harnessing the power of multi-task pretraining for ground-truth level natural language explanations. *CoRR*, abs/2212.04231.

Yue Qiu, Yoshiki Nagasaki, Kensho Hara, Hirokatsu Kataoka, Ryota Suzuki, Kenji Iwata, and Yutaka Satoh. 2023. Virtualhome action genome: A simulated spatio-temporal scene graph dataset with consistent relationship labels. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2023, Waikoloa, HI, USA, January 2-7, 2023*, pages 3340–3349. IEEE.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 91–99.

Fawaz Sammani, Tanmoy Mukherjee, and Nikos Deligiannis. 2022. NLX-GPT: A model for natural language explanations in vision and vision-language tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 8312–8322. IEEE.

Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. 2020a. Long-tailed classification by keeping the good and removing the bad momentum causal effect. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. 2020b. Unbiased scene graph generation from biased training. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 3713–3722. Computer Vision Foundation / IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4566–4575. IEEE Computer Society.


(c) 2024 中国中文信息学会计算语言学专业委员会

Tzu-Jui Julius Wang, Selen Pehlivan, and Jorma Laaksonen. 2020. Tackling the unannotated: Scene graph generation with bias-reduced models. In *31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020*. BMVA Press.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022a. OFA: unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 23318–23340. PMLR.

Zhecan Wang, Haoxuan You, Liunian Harold Li, Alireza Zareian, Suji Park, Yiqing Liang, Kai-Wei Chang, and Shih-Fu Chang. 2022b. SGEITL: scene graph enhanced image-text learning for visual commonsense reasoning. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 5914–5922. AAAI Press.

Chenxi Whitehouse, Tillman Weyde, and Pranava Madhyastha. 2023. Towards a unified model for generating answers and explanations in visual question answering. In Andreas Vlachos and Isabelle Augenstein, editors, *Findings of the Association for Computational Linguistics: EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 1648–1660. Association for Computational Linguistics.

Danfei Xu, Yuke Zhu, Christopher B. Choy, and Li Fei-Fei. 2017. Scene graph generation by iterative message passing. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3097–3106. IEEE Computer Society.

Shaotian Yan, Chen Shen, Zhongming Jin, Jianqiang Huang, Rongxin Jiang, Yaowu Chen, and Xian-Sheng Hua. 2020. PCPL: predicate-correlation perception learning for unbiased scene graph generation. In Chang Wen Chen, Rita Cucchiara, Xian-Sheng Hua, Guo-Jun Qi, Elisa Ricci, Zhengyou Zhang, and Roger Zimmermann, editors, *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pages 265–273. ACM.

Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 3208–3216. AAAI Press.

Alireza Zareian, Svebor Karaman, and Shih-Fu Chang. 2020. Weakly supervised visual semantic parsing. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 3733–3742. Computer Vision Foundation / IEEE.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6720–6731. Computer Vision Foundation / IEEE.

Hanwang Zhang, Zawlin Kyaw, Jinyang Yu, and Shih-Fu Chang. 2017. PPR-FCN: weakly supervised visual relation detection via parallel pairwise R-FCN. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 4243–4251. IEEE Computer Society.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal chain-of-thought reasoning in language models. *CoRR*, abs/2302.00923.

CCL 2024

# 基于逻辑推理和多任务融合的认知刺激对话生成方法

**蒋玉茹[1]，李梦媛[1]*，陶宇阳[1]，区可明[2]，佘泽鹏[2]，施水才[1,2]**
1. 北京信息科技大学计算机学院，北京
2. 北京拓尔思信息技术有限公司，北京
{jiangyuru,limengyuan}@bistu.edu.cn
{ou.keming,she.zepeng,shi.shuicai}@trs.com.cn

## 摘要

在全球老龄化背景下，带有认知刺激的对话系统是保持老年人认知健康的重要手段。中文认知刺激对话数据集(Chinese Cognitive Stimulation Conversation Dataset, CSConv)和模型构建的研究工作刚刚开始。本文将认知刺激对话生成视为一个多任务融合的逻辑思维推理过程，将情感分类任务、决策任务和对话回复生成任务间的逻辑关系，建模为一个推理过程，来引导大语言模型生成。针对决策任务，本文提出分层编码器结构的决策模型。决策实验结果表明，决策模型有效的提高了决策任务的准确率。针对多任务过程，本文提出多任务融合方法，将三个任务对应的模型结合在一起。生成实验结果表明，分类、决策及生成的多任务融合方法，显著提升了对话回复能力，证明了该方法的有效性和先进性。

**关键词：** 认知刺激对话；逻辑推理；决策任务；多任务融合方法

# Cognitive stimulation dialogue generation method based on logical reasoning and multi-task integration

**Yuru Jiang[1]，Mengyuan Li[1]*，Yuyang Tao[1]，Keming Ou[2]，Zepeng She[2]，Shuicai Shi[1,2]**
1. School of Computer Science,
Beijing Information Science and Technology University
2. TRS Information Technology Co.,Ltd.，Beijing
{jiangyuru,limengyuan}@bistu.edu.cn
{ou.keming,she.zepeng,shi.shuicai}@trs.com.cn

## Abstract

In the context of global aging, dialogue systems with cognitive stimulation are an important means to maintain the cognitive health of the elderly. Research work on the Chinese Cognitive Stimulation Conversation dataset (CSConv) and model construction has just begun. This study regards cognitive stimulation dialogue generation as a multi-task integration thinking and reasoning process, and models the logical relationship between emotion classification tasks, decision-making tasks and dialogue response generation tasks as a reasoning process to guide the generation of large language models. For decision-making tasks, this paper proposes a decision-making model with a hierarchical encoder structure. The decision-making experiment results show that the decision-making model effectively improves the accuracy of decision-making tasks. For the multi-task process, this study proposes a multi-task integration method to combine

* 通讯作者
©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版
**基金项目：**北京市自然科学基金项目（4242019），题目：智慧养老服务中对话机器人关键技术研究

the models corresponding to the three tasks. The generation experimental results show that the multi-task integration method of classification, decision-making and generation significantly improves the dialogue response ability, proving the effectiveness and advancement of the method.

**Keywords:** Cognitively stimulating dialogue , Logical reasoning , Decision-making tasks , Multi-task integration method

## 1 引言

人机对话系统一直是人工智能研究中的热点。在全球老龄化背景下，利用人工智能实现智慧康养，构建面向老人的对话系统成为对话系统研究的重要方向。研究表明，对话系统对老人认知能力的改善具有积极的作用(De Oliveira et al., 2014; Park et al., 2019)。其中，认知刺激对话是恢复老人认知功能的重要心理治疗手段(Tokunaga et al., 2019; Tokunaga et al., 2021)。认知刺激治疗是一种帮助轻度到重度以下认知障碍症患者的非药物干预治疗手段(Spector et al., 2003)。认知刺激对话是在认知刺激治疗原则[1]指导下，以刺激患者思维为目的生成的对话。结合认知刺激治疗的生成式对话系统在智慧养老领域中有着广阔的应用场景，可以应用于家居生活场景和养老社区服务场景等，通过聊天的方式，帮助老人认知恢复。因此认知刺激对话有着迫切的现实需求、应用价值和多样的应用场景。

之前大部分面向老年人的对话系统主要关注在特定场景中的特定功能研究，例如Tokunaga et al. (2021)等人，提出的一个利用照片故事的机器人对话系统，对老人进行日常的认知训练。Navarro et al. (2018)等人提出了一个认知刺激的治疗系统，用以评估患者的认知障碍和情绪健康，来自动生成新的个性化刺激计划。然而此前，由于认知刺激对话语料的缺失，没有面向关于认知刺激对话的生成式系统研究。

近期，面向老人的生成式认知刺激对话系统研究取得了突破。Jiang et al. (2023)构建了一个中文认知刺激对话数据集CSConv，将情感支持(Liu et al., 2021)和认知刺激同时引入聊天对话中，提出了面向老年人的认知刺激对话新任务。数据集包含2600个对话，每句话语标注了三个标签：认知刺激治疗原则标签、情感标签和情感支持策略标签，其中情感支持策略标签由Liu et al. (2021)提出，认知刺激治疗原则标签由Jiang et al. (2023)从认知刺激治疗手册中总结而来。然后，作者提出了融合多源知识的生成式对话模型CSD，取得了最好的效果。




图 1: 人类思维流程图

认知刺激对话任务不同于普通的闲聊任务，它是带有目的性的对话，任务难度比较高。在进行认知刺激对话的时候，人类思维首先会根据当前话语的内容进行决策，确定适用的治疗原则，随后根据决策的结果进行相应的回应，如图1所示。在情感支持对话(Liu et al., 2021)和认知刺激对话(Jiang et al., 2023)工作中，都提供了一个对话示例，示例中的灰色框描述了在输入和回复之间回复者的思考过程。由此可见，模型在回复时同样会参考人类思维的过程，即先决策后回复。这也表明，认知刺激对话任务是多任务的融合。

因此，本文提出一个多任务融合的方法，如图2所示，把认知刺激对话任务分成三个子任务，即分类任务、决策任务和对话回复生成任务。本文将三个任务间的逻辑关系，视作一个复杂的逻辑推理过程，来引导大语言模型生成。分类任务即情感信息分类，决策任务具体分为认知刺激治疗决策和情感支持策略决策，生成任务为生成模型利用情感信息和决策信息的引导来做出相应的回复。

[1]《认知刺激治疗CST：为认知障碍症设计的循证小组活动（导师手册）》


(c) 2024 中国中文信息学会计算语言学专业委员会

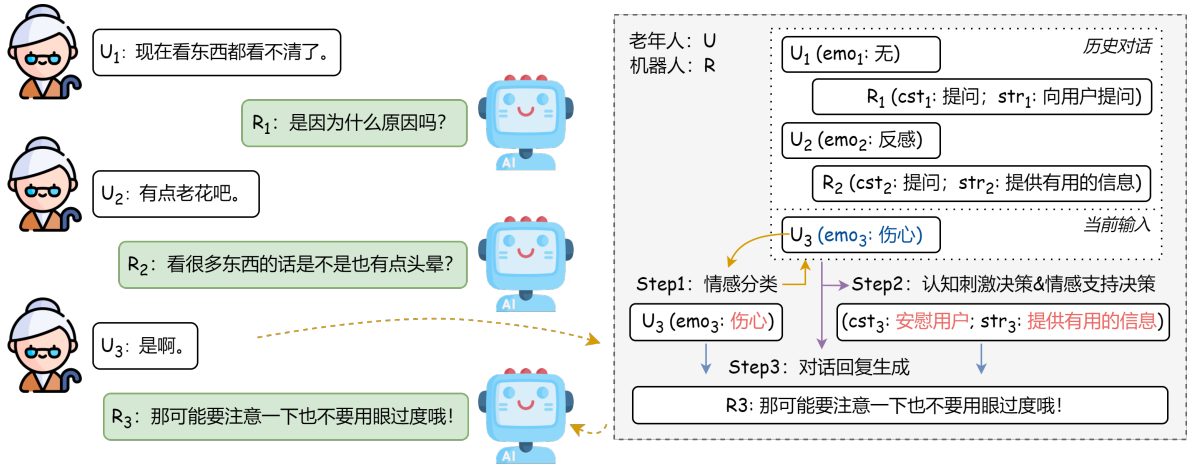



图 2: 多任务融合流程图

针对决策任务，本文提出分层编码器结构的决策模型。在编码阶段，通过融入历史对话，情感信息以及历史决策信息，来增强对话的编码表示，从而提升决策准确率，为生成模型提供更准确的信息。对于生成模型而言，从情感信息和决策信息到回复生成中间还有很大的跨度。生成模型不仅要理解标签含义，还肩负着理解情感信息和决策信息背后蕴藏的指令。大语言模型自出现以来，展现了惊人的推理能力和对指令的服从能力，可以很好的理解指令信息并做出相应的回应，对于不同标签信息背后的含义有着更好的理解。于是本研究使用大语言模型作为生成模型，使用基于提示的方法，探究大语言模型深层次的理解和推理能力。

本研究具体的研究贡献总结如下：

（1）针对认知刺激对话任务，本文参考人类先决策后回复的思维过程，提出多任务流程。将认知刺激对话分为三个子任务，具体为情感分类任务，决策任务，以及大语言模型对话回复生成任务。

（2）针对决策任务，本文提出分层编码器结构的决策模型DM。决策实验结果表明，DM有效的提高了情感支持决策和认知刺激决策的准确率。生成实验表明，随着决策模型准确率的提高，对生成模型生成对话回复也具有正向的促进作用。决策实验和生成实验的结果证明了决策任务和决策模型的有效性。

（3）针对认知刺激对话过程中多任务的特点，本文提出了多任务融合方法CSMI，将情感分类模型、决策模型和大语言模型融合到一起。本研究进行了大量的实验，结果表明CSMI提升了对话回复质量，证明了CSMI的有效性与先进性。

## 2 相关工作

在面向老人的对话系统研究中，一些研究表明人机对话系统可以帮助老人应对孤独，降低老人的抑郁和焦虑水平，改善生活质量。Ryu et al. (2020)设计和开发了一个心理健康护理对话机器人Yeonheebot，来减少老人的焦虑和抑郁。Yeonheebot会启动关于老年人幸福日常护理的对话，并提供与他们兴趣相关的功能，以吸引用户使用。名为Charlie(Valtolina and Hu, 2021)的面向老人的陪伴型聊天机器人，能提供智力游戏、能主动提醒事务、能聊一些轶事以促进自我共情，意图帮助老人积极应对孤独感，改善生活质量。Lee et al. (2019)也对在对话中融入自我共情(Sharma et al., 2021; Sharma et al., 2020; Rashkin et al., 2019)进行了研究，其实验表明：让用户去关心机器人比让机器人关心用户更能让用户获得自我共情。

还有一些研究表明人机对话系统对老人的认知能力改善具有积极的作用。Tokunaga et al. (2021)研究了一个名字叫Bono 的对话系统，提出一个基于对话的方法对老人进行日常的认知训练，利用对话机器人，向老人展示照片并讲述相关故事，以此作为和老人聊天的上下文，然后请老人提问，机器人进行回答。Kim et al. (2021)利用基于智能语音程序的元记忆训练策略，进行了实验，实验结果表明该训练方法能够帮助到具有主观认知能力下降的老人。

De Oliveira et al. (2014)研究表明多感官认知刺激对长期护理机构中老人的认知能力问题具有有益影响。Park et al. (2019)研究了多组件的认知刺激程序对老年人的认知功能改善具有积


(c) 2024 中国中文信息学会计算语言学专业委员会

极的作用。Navarro et al. (2018)提出一个认知刺激治疗系统，用以评估患者的认知障碍和情绪健康，来自动生成新的个性化刺激计划。可见，认知刺激作为针对认知障碍老人的非药物心理治疗方法，同样对老人的认知能力改善具有重要作用。

近期，认知刺激和生成式对话系统结合的研究上取得了进展，Jiang et al. (2023)构建了一个中文认知刺激对话数据集CSConv，首次提出面向老年人的认知刺激对话任务。其中包含2600个对话，每句话语标注了三个类别标签，然后提出了融合多源知识的生成式对话模型，取得了当前最好的效果。

## 3 多任务融合的认知刺激对话方法

本文参考人类在认知刺激对话中，分步进行的特点，提出了多任务过程，并基于此设计了多任务融合的方法，以下简称为CSMI（Cognitively Stimulated dialogue based on Multi-task Integration）。本研究将认知刺激对话任务分成三个子任务，分别是分类任务、决策任务和大语言模型对话回复生成任务。

具体来说，一本文选用BERT(Devlin et al., 2018)作为情感分类模型，针对情感分类任务构建了相应的数据集，并利用该数据集对模型进行微调。二本文构建了决策模型，其包括两个主要阶段：（1）对话编码和（2）标签决策。在标签决策阶段，本研究设计了两种结构，分别是基于LSTM(Shi et al., 2015)构建的决策模型，以下简称$\mathrm{DM}_{LSTM}$（Decision Model-LSTM），和基于Transformer Encoder(Vaswani et al., 2017)构建的决策模型，以下简称$\mathrm{DM}_{Transformer}$（Decision Model-Transformer）。三本文采用提示的方法，进一步提升大语言模型的理解能力和推理能力。最后，结合分类模型和决策模型的结果，将对应三个任务的模型融合在一起，来引导大语言模型生成回复。

### 3.1 数据集处理

在CSConv数据集中，为每句话语$u$都标注了三种类型的标签：情感标签$u_{emo}$、认知刺激治疗原则标签$u_{cst}$、情感支持策略标签$u_{strategy}$。在Jiang et al. (2023)工作中，作者把每一个标签转换成一个词进行编码，拼接在句子编码之后输入给生成模型，因此没有给出相应的中文标签。于是，对于情感标签和认知刺激治疗原则标签，本研究按照Jiang et al. (2023)工作中的释义，翻译成中文后，再总结为对应的中文标签，如表1所示。对于情感支持策略标签，按照Liu et al. (2021)给出的示例，总结成对应的中文标签，如表2所示。

| 英文标签 | 中文标签 |
|---|---|
| None | 无 |
| Disgust | 反感 |
| Sadness | 伤心 |
| Fear | 害怕 |
| Surprise | 惊喜 |
| Like | 喜欢 |
| Happiness | 幸福 |
| Anger | 愤怒 |

(a) 情感标签对照表

| 英文标签 | 中文标签 |
|---|---|
| None | 无 |
| Inquiry | 提问 |
| Respect | 尊重 |
| Reminiscence | 回忆过去 |
| Expression | 提高用户语言表达能力 |
| Enjoyment | 在谈话中获得乐趣 |
| Comfort | 安慰用户 |

(b) 认知刺激治疗原则标签对照表

表 1: 情感和认知刺激治疗原则中文标签对照表

决策任务应预测出机器回复要采取的情感支持策略或者认知刺激治疗原则。于是根据决策任务的定义，本文重新整理了CSConv数据集。在老人话语句$u$留下情感信息标签$u_{emo}$，在机器回复句$r$留下情感支持策略标签$r_{strategy}$和认知刺激治疗原则标签$r_{cst}$。

### 3.2 多任务融合方法

针对决策任务，本文提出具有分层编码器结构的决策模型DM，包含两个阶段：（1）对话编码（2）标签决策。在标签决策阶段，本文通过两种不同的结构设计了决策模

| 英文标签 | 中文标签 |
|---|---|
| None | 无 |
| Question | 向用户提问 |
| Restatement or Paraphrasing | 进行更简洁的重新表述 |
| Reflection of Feelings | 描述用户感受 |
| Self-disclosure | 利用类似经历或情感，表达同理心 |
| Affirmation and Reassurance | 肯定优势和能力，提供鼓励 |
| Providing Suggestions | 提供改变的建议 |
| Information | 提供有用的信息 |
| Others | 其他 |

表 2: 情感支持策略中文标签对照表

型$\mathrm{DM}_{LSTM}$和$\mathrm{DM}_{Transformer}$。决策模型结构如图3所示，图片使用认知刺激决策为例进行展示，情感支持决策与此类似。

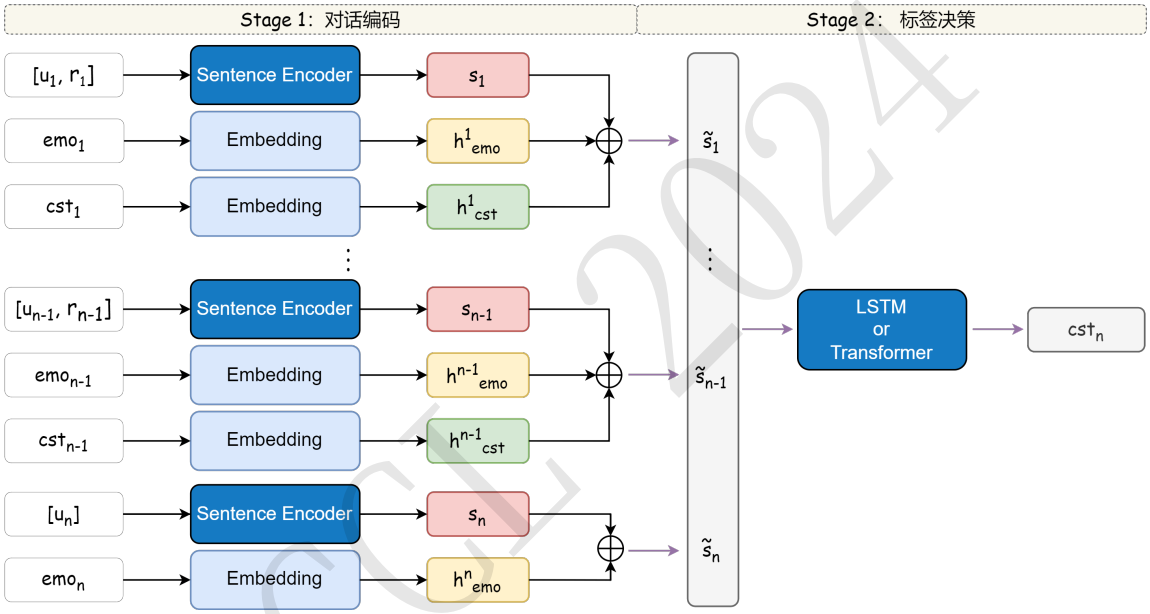



图 3: 决策模型结构图

本文将对话历史中的每一轮设计成一个句对的形式$H = [(u_1, r_1), ..., (u_{n-1}, r_{n-1})]$进行编码，以提高历史对话的结构化表达，其中$u$代表老人的话语，$r$代表机器人的话语。为了更有效地传达句子的语义信息，本文采用句子编码器Sentence BERT(Reimers and Gurevych, 2019)对它们进行编码。再和当前句$u_n$一起，通过句子编码器做句子表征，如公式（1）所示。若设$s = (u, r)$，则得到对话表征$S = [s_1, s_2, ..., s_n]$，具体表示为：

$$S = SentenceBERT([(u_1, r_1), ..., (u_{n-1}, r_{n-1}), u_n]), S \in \mathbb{R}^{n \times d} \quad (1)$$

其中，$n$表示对话轮数，$d$为句子编码器隐含层维度。然后，本文使用两种编码层分别对每一轮的情感信息和历史对话中的决策信息进行编码，即$E_e(emo) \in \mathbb{R}^{n \times d}$、$E_c(cst) \in \mathbb{R}^{n \times d}$或$E_s(str) \in \mathbb{R}^{n \times d}$，其中$n$表示对话轮数，$d$为隐藏层维度。我们将历史情感信息和历史决策信息组成序列，经过编码层后得到输出结果：$H_{emo}$、$H_{cst}$和$H_{str}$，$h_p^n$表示占位符的编码层结


(c) 2024 中国中文信息学会计算语言学专业委员会

果。

$$H_{emo} = (h^1_{emo}, h^2_{emo}, ..., h^n_{emo}) \tag{2}$$
$$H_{cst} = (h^1_{cst}, h^2_{cst}, ..., h^{n-1}_{cst}, h^n_p) \tag{3}$$
$$H_{str} = (h^1_{str}, h^2_{str}, ..., h^{n-1}_{str}, h^n_p) \tag{4}$$

将编码层得到的输出结果，与对应的每一轮对话编码结果进行拼接，得到信息增强后的对话编码结果：融合情感和认知刺激决策信息的对话编码$\widetilde{S}_{cst}$以及融合情感和情感支持决策信息的对话编码$\widetilde{S}_{str}$。

$$\widetilde{S}_{cst} = S \oplus H_{emo} \oplus H_{cst} \tag{5}$$
$$\widetilde{S}_{str} = S \oplus H_{emo} \oplus H_{str} \tag{6}$$

对话是一种具有前后顺序关系和上下文联系的序列，因此，在本研究的标签决策阶段中，采用了长短期记忆网络（LSTM）和Transformer Encoder两种不同结构。

LSTM被视为一种具有时间序列特征的递归神经网络模型，其在模型训练中能够有效地捕获句对之间的时序关联，从而有助于深入理解对话中的上下文语境。将对话编码结果传入LSTM得到结果$G_{cst}$和$G_{str}$，取其中最后一层的隐含层向量$g^n_{cst}$和$g^n_{str}$，经过线性层后得到当前句的决策信息，$W_l$为决策预测权重矩阵：

$$G_{cst} = LSTM(\widetilde{S}_{cst}) \tag{7}$$
$$l_{cst} = argmax(softmax(W_l g^n_{cst})) \tag{8}$$
$$G_{str} = LSTM(\widetilde{S}_{str}) \tag{9}$$
$$l_{str} = argmax(softmax(W_l g^n_{str})) \tag{10}$$

Transformer Encoder是一种基于自注意力机制的模型，能够通过自注意力有效地建模句子中词与词之间的关系，从而实现对对话信息的深入理解。为了促进Transformer Encoder对句与句关系更为深入的理解，本文采用一个位置编码层$E_p(hp) \in \mathbb{R}^{n \times d}$表示句对之间的上下文关系，其中$n$表示对话轮数，$d$为隐藏层维度与句子编码器隐藏层向量维度保持一致。将对话编码结果和位置编码结果相加，传入Transformer Encoder得到结果$M_{cst}$和$M_{str}$，取其中最后一层的隐含层向量$m^n_{cst}$和$m^n_{str}$，经过线性层后得到当前句的决策信息，$W_l$为决策预测权重矩阵：

$$H_p = (h^1_p, h^2_p, ..., h^n_p) \tag{11}$$
$$M_{cst} = TransformerEncoder(\widetilde{S}_{cst} + H_p) \tag{12}$$
$$l_{cst} = argmax(softmax(W_l m^n_{cst})) \tag{13}$$
$$M_{str} = TransformerEncoder(\widetilde{S}_{str} + H_p) \tag{14}$$
$$l_{str} = argmax(softmax(W_l m^n_{str})) \tag{15}$$

针对大语言模型，本文选择中文大语言模型ChatGLM(Du et al., 2022)作为生成模型，并采用基于提示的方法，设计了清晰高效的提示模板，将任务解释和逻辑过程设计成提示模板，嵌入在生成模型的对话模板中。任务提示模板嵌入在ChatGLM对话模板中的“问：”之后，拼接在每次对话的开始，旨在更有效地引导大语言模型了解决策任务，增进其对标签语义信息的理解。任务提示模板设计如下所示：“你是一个具备认知刺激和情感支持功能的聊天机器人，你的聊天对象是一位老人，你要根据认知刺激策略和情感支持策略选择恰当的回复。现在开始聊天：{用户输入}”。对于回复提示模板，本文将三个任务间的逻辑关系，即“根据用户情感和发言决定回复”，视作一个复杂的逻辑思维过程，以此设计提示来引导大语言模型生成。回复模板设计如下：“根据用户发言中的情绪为[{情感标签}]，应该选择的认知刺激疗法是[{原则标签}]，使用的情感支持策略是[{策略标签}]，所以回复为”。将回复提示模板嵌入到ChatGLM对话模板中的“答：”之后。随后微调大语言模型，在微调时，模板中的占位符将被替换为具体的标签信息。

CCL 2024


(c) 2024 中国中文信息学会计算语言学专业委员会

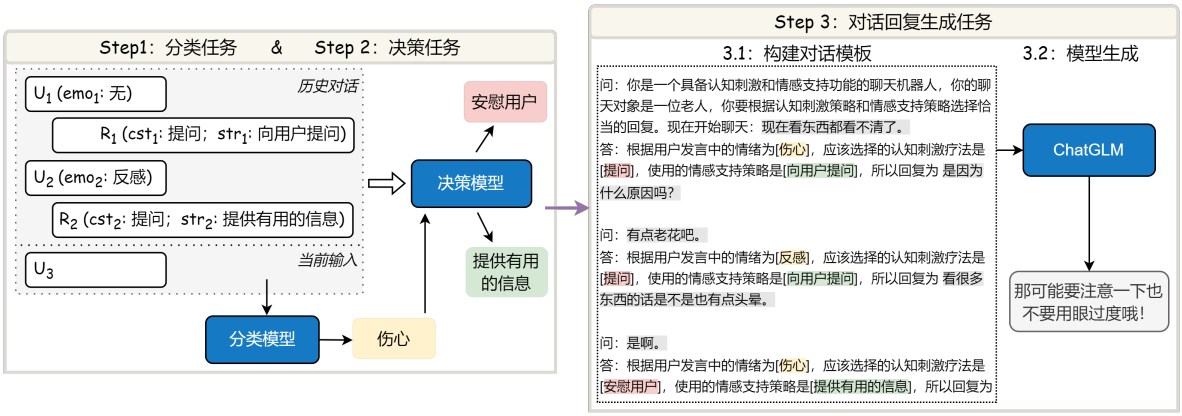



图 4: 多任务融合方法总体模型图

多任务融合方法模型图如图 4所示。首先，将当前话语通过情感分类得到情感信息。然后，结合当前话语及其历史对话和情感信息通过决策模型得到认知刺激治疗原则与情感支持策略两个决策信息。最后，结合情感分类模型和决策模型的结果，将所有需要的信息嵌入在对话模板中，通过生成模型得到对话回复。

## 4 实验与分析

### 4.1 实验设置

在情感分类模型中，我们使用的预训练模型是bert-base-Chinese，设置学习率为2e-5。在决策模型中，句子编码器使用的预训练模型是sbert-base-chinese-nli，设置学习率为2e-5，LSTM和Transformer Encoder的学习率为1e-4，批次大小为8，累积步数是4。情感信息和决策信息编码层的向量维度设置为32，输入的最大历史对话长度设置为3轮对话。对于生成模型，我们设计了任务提示模板和回复提示模板，将其添加到ChatGLM的回复模板中，并使用Ptuning方法微调ChatGLM[2]。设置最大序列长度为512，批次大小为16，累积步数为2。

### 4.2 评价指标

在决策模型的评估中，本研究选用准确率（Accuracy）和F1分数（F1-score）作为主要评价指标。准确率旨在度量模型正确预测标签的数量，而F1分数则综合考虑了精确率和召回率，为模型的综合性能提供了评估标准。

本研究沿用前人的工作，以BLEU(Papineni et al., 2002)作为评估对话生成质量的指标，BLEU着眼于衡量生成文本与参考文本之间的语义相似性。还引入了ROUGE(Lin, 2004)来评估对话生成质量，ROUGE侧重于评估生成文本对信息的涵盖度和完整性。通常情况下，这两种指标结合使用，以全面评估模型的性能。此外，本研究还统计了BERTScore(Zhang et al., 2019)，用以度量生成文本与参考文本之间的语义相似性，进一步丰富了评价的维度。

### 4.3 对比实验

为了全面评估本文提出的决策模型在决策任务上的表现，本研究设计了一系列对比实验，涉及多种深度学习模型，包含基于LSTM构建的决策模型$\mathrm{DM}_{LSTM}$和基于Transformer Encoder构建的决策模型$\mathrm{DM}_{Transformer}$。此外，为了确保评估的全面性和深度，本研究还纳入了大语言模型作为决策模型的基线方法。由于大语言模型具有推理能力，所以本文尝试让大语言模型自己担任决策模型的角色。思维链（CoT）(Wei et al., 2022)作为一种设计Prompt(Liu et al., 2023)的方法，即Prompt中除了有任务的输入和输出外，还包含推理的中间步骤。由于认知刺激对话任务，从输入到输出也存在着逻辑关系，所以本文选择使用CoT的方法，让大语言模型自己作为决策模型，思维链的设计参考3.2中的回复提示模板。

如表3所示，$\mathrm{DM}_{Transformer}$和$\mathrm{DM}_{LSTM}$在两个决策任务的准确率和F1分数上高于其他基线，由此证明了两种决策模型$\mathrm{DM}_{Transformer}$和$\mathrm{DM}_{LSTM}$具备有效性与先进性。

[2]https://github.com/THUDM/ChatGLM2-6B


(c) 2024 中国中文信息学会计算语言学专业委员会

| 模型名称 | 认知刺激决策 | | 情感支持决策 | |
|---|---|---|---|---|
| | acc | f1 | acc | f1 |
| 思维链推理 | 62.81 | 62.94 | 65.96 | 65.29 |
| BERT | 63.41 | 63.38 | 70.53 | 70.16 |
| **DM$_{Transformer}$** | **65.96** | **66.34** | **72.54** | **70.93** |
| **DM$_{LSTM}$** | **67.37** | **67.14** | **72.63** | **71.92** |

表 3: 决策任务对比实验结果

| 模型名称 | BLEU-2 | BLEU-4 | BERTScore | ROUGE-1 | ROUGE-2 | ROUGE-L |
|---|---|---|---|---|---|---|
| CDialGPT$_{base}$ | 17.55 | 6.22 | 57.70 | - | - | - |
| CDialGPT$_{large}$ | 15.05 | 5.47 | 57.81 | - | - | - |
| GPT2-chitchat | 34.61 | 21.04 | 66.37 | - | - | - |
| Distil-cluecorpussmall | 39.94 | 25.30 | 69.41 | - | - | - |
| Cluecorpussmall | 41.04 | 26.59 | 68.65 | - | - | - |
| CSD | 45.53 | 30.90 | 74.61 | - | - | - |
| **ChatGLM** | **46.30** | **39.32** | **77.67** | **49.89** | **36.82** | **49.51** |
| **CLC** | **43.86** | **36.27** | **77.08** | **46.91** | **33.80** | **46.28** |
| **CSMI-BERT** | **44.35** | **36.91** | **77.13** | **48.27** | **34.02** | **47.42** |
| **CSMI-DM$_{Transformer}$** | **47.92** | **37.23** | **77.12** | **48.18** | **33.94** | **47.57** |
| **CSMI-DM$_{LSTM}$** | **48.83** | **38.85** | **78.27** | **50.34** | **36.27** | **49.71** |
| **CSMI-真实标签** | **50.41** | **41.26** | **79.85** | **54.87** | **38.56** | **54.19** |

表 4: 生成任务对比实验结果

对于对话回复生成任务，以上述分类模型、决策模型和大语言模型ChatGLM为骨干，本文构建了以下变体模型：（1）微调版本（ChatGLM）：直接使用CSConv数据集微调ChatGLM。（2）基于思维链微调（CLC）：使用思维链方法微调大语言模型。（3）多任务融合方法（CSMI）：使用基于提示的方法微调大语言模型。此外，本文亦通过对CSConv数据集上的先前研究成果进行对比，来体现本研究方法的先进性。包括：在LCCC不同版本数据集上训练的CDialGPT$_{base}$(Wang et al., 2020)和CDialGPT$_{large}$(Wang et al., 2020)；在闲聊语料上训练的GPT2-chitchat；在CLUECorpusSmall语料库上训练的Distil-Cluecorpussmall(Radford et al., 2019)和Cluecorpussmall(Radford et al., 2019)，以及Jiang et al. (2023)提出的中文认知刺激对话系统CSD。

如表4所示，CSMI方法的评分高于前人的工作，由此证明多任务融合方法的有效性与先进性。同时，CSMI与不同决策模型的评分随着决策准确率的提升而提高，表明决策信息对于生成模型有着正向的引导作用，证明了多任务融合方法的有效性与先进性。但是，CSMI与不同决策模型的评分略低于ChatGLM和CSMI-真实标签的评分，因为如果决策策略与真实值不同，则生成的回复与真实回复会有很大的差异。由此可见，在没有提供真实标签的情况下，学习决策信息是很重要的。

### 4.4 消融实验

为了进一步验证决策模型和提示模板在大语言模型上的表现，本文设计了一系列消融实验。

本文通过对话编码中是否融入情感信息和决策信息作为控制条件进行实验，表5展示了以LSTM构成的决策模型DM$_{LSTM}$，针对每一部分设计，在认知刺激决策任务上的消融实验结果。可以看到，历史对话信息、情感信息和历史决策信息，都对决策模型DM$_{LSTM}$产生贡献。其中情感信息和历史决策信息的贡献更为突出，表明模型可以有效的利用情感和决策信息，进


(c) 2024 中国中文信息学会计算语言学专业委员会

| 模型名称 | 历史轮数 | 含情感信息 | 含历史决策信息 | ACC | F1 |
|---|---|---|---|---|---|
| $\text{DM}_{LSTM}$ | 1 | 否 | 否 | 63.51 | 63.79 |
| $\text{DM}_{LSTM}$ | 3 | 否 | 否 | 65.26 | 65.65 |
| $\text{DM}_{LSTM}$ | 3 | 是 | 否 | 66.67 | 66.32 |
| **$\text{DM}_{LSTM}$** | **3** | **是** | **是** | **67.37** | **67.14** |

表 5: $\text{DM}_{LSTM}$模型在认知刺激决策任务上的消融实验结果

一步提升模型决策能力。

本文还通过控制对大语言模型不同的提示模板信息为条件进行实验。如表6所示，加入了回复提示模板微调的模型优于直接微调的模型，可见把三个任务间的推理过程作为回复提示，进一步提高了大语言模型对于标签信息的理解能力。在增加了任务提示模板后，BLEU-4和ROUGE-L分数均又提高了1.5左右，由此证明基于提示的方法可以激发大语言模型对任务深层次的理解和推理能力，表明了基于提示方法的有效性。

| 模型名称 | 回复提示模板 | 任务提示模板 | BLEU-4 | ROUGE-L |
|---|---|---|---|---|
| ChatGLM | 否 | 否 | 39.32 | 49.51 |
| CSMI | 是 | 否 | 39.62 | 52.62 |
| **CSMI** | **是** | **是** | **41.26** | **54.19** |

表 6: 基于提示模板的大语言模型消融实验

### 4.5 人工评测

对话生成任务中，自动化评测指标往往存在局限性，因为其无法反映出模型所生成的结果与参考目标之间在深层语义及流畅度等方面的差距。因此本文开展了人工评测，从下述三个主要评价角度衡量模型的生成能力：

- 流畅度：机器回复是否流畅并符合中文语法。
- 支持性：模型回复能否提供帮助。
- 决策匹配度：机器回复是否能根据给出的决策信息，生成对应的回复内容。

| 对话主题 | 流畅度 | 支持性 | 决策匹配度 | |
|---|---|---|---|---|
| | | | 认知刺激 | 情感支持 |
| 出题游戏 | 4.61 | 3.64 | 4.24 | 4.07 |
| 回忆过往 | 4.28 | 4.06 | 4.67 | 4.39 |
| 分享生活 | 4.43 | 3.98 | 4.41 | 4.56 |
| **平均分** | **4.44** | **3.89** | **4.62** | **4.34** |

表 7: 人工评测结果

本文选取CSMI-$\text{DM}_{LSTM}$作为人工评测的模型对象，保留其在测试集上的生成结果。本研究从测试集中选取了三个出现频次较多的对话主题作为评测内容，分别是出题游戏、回忆过往和分享生活，共三段完整的多轮对话。以上评测标准的衡量标准均为1-5分，分别表示不可接受-优异。本研究邀请5名评测人员参加人工评测，结果如表7所示。总体来看，流畅度和决策匹配度都达到了良好的水平。支持性评分表明，评测人员认为机器在对话中给出的回复可以提供些许帮助。决策匹配度评分相对较高，说明决策信息作为一种指令，可以很好的引导大语言模型生成回复。


(c) 2024 中国中文信息学会计算语言学专业委员会

### 4.6 样例分析

| | |
|---|---|
| 历史对话 | $U_1$：我昨天做了个小手工。<br>$R_1$：做了个什么呢?<br>$U_2$：拿了干花粘在了木框上做成了相框。<br>$R_2$：很有意思诶。 |
| 当前输入 | $U_3$：但是我不会搭配颜色，粘的很乱。 |
| 情感分类结果 | emo：伤心 |
| 决策结果 | cst：安慰用户；str：肯定优势和能力，提供鼓励 |
| **模型回复** | **$R_3$：没有啊！我觉得很有创意，真是太厉害了。** |

表 8：样例分析

本文从测试集生成的结果中选取了一个对话样本作为分析样例，同时保存了其生成时情感分类以及决策的中间结果，如表8所示。老人分享了自己做手工的事情，但是觉得自己做的不完美，给出了“但是我不会搭配颜色，粘的很乱。”的回复。对于老人的这一回答，情感分类模型对于当前句分类为伤心，决策模型给出了安慰用户以及提供鼓励的决策结果，生成模型根据以上结果给出回复“没有啊！我觉得很有创意，真是太厉害了。”。模型的回复在安慰老人的同时又给出了自己认为很有创意的肯定和鼓励。由此可见，情感信息和决策信息可以作为一种指令，来引导大语言模型生成与它们高度相关的回复。同时也表明了情感和决策判断的正确与否，会对后续生成产生影响。表明了学习情感信息和决策信息的重要性，证明了多任务融合方法的有效性。

## 5 总结

本文将认知刺激对话任务视为具有分类、决策和生成任务的多任务融合推理过程。重新整理了决策数据集，设计了带有分层编码器的决策模型。并将情感分类模型、决策模型和生成模型融合到一起，提出多任务融合方法。决策实验以及生成实验，都证明了本研究方法的有效性和先进性。本文研究的这一进展，为认知刺激对话任务提供了新的视角，不仅拓展了对认知刺激对话任务的理解，而且为相关领域的未来研究提供了有价值的参考和启示。同时也为认知刺激对话在结合大型语言模型方面开辟了新的方向。由于我们方法灵活多样，未来将着重设计更加强大的决策模型，并且探索结合不同的大语言模型。这些努力将为中文认知刺激对话任务的发展做出新的贡献。

## 参考文献

T. C. De Oliveira, F. C. Soares, L. D. De Macedo, D. L. Diniz, N. V. Bento-Torres, and C. W. o Diniz. 2014. Beneficial effects of multisensory and cognitive stimulation on age-related cognitive decline in long-term-care institutions. *Clin Interv Aging*, 9:309–320.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Jiyue Jiang, Sheng Wang, Qintong Li, Lingpeng Kong, and Chuan Wu. 2023. A cognitive stimulation dialogue system with multi-source knowledge fusion for elders with cognitive impairment. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10628–10640, Toronto, Canada, July. Association for Computational Linguistics.


(c) 2024 中国中文信息学会计算语言学专业委员会

Jeongsim Kim, EunJi Shin, KyungHwa Han, Soowon Park, Jung Hae Youn, Guixiang Jin, and Jun-Young Lee. 2021. Efficacy of smart speaker–based metamemory training in older adults: Case-control cohort study. *Journal of medical Internet research*, 23(2):e20177.

Minha Lee, Sander Ackermans, Nena Van As, Hanwen Chang, Enzo Lucas, and Wijnand IJsselsteijn. 2019. Caring for vincent: a chatbot for self-compassion. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. *arXiv preprint arXiv:2106.01144*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Javier Navarro, Faiyaz Doctor, Víctor Zamudio, Rahat Iqbal, Arun Kumar Sangaiah, and Carlos Lino. 2018. Fuzzy adaptive cognitive stimulation therapy generation for alzheimer's sufferers: Towards a pervasive dementia care monitoring platform. *Future Generation Computer Systems*, 88:479–490.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Jeong-Mo Park, Mi-Won Kim, and Hee-Young Shim. 2019. Effects of a multicomponent cognitive stimulation program on cognitive function improvement among elderly women. *Asian Nursing Research*, 13(5):306–312.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy, July. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Hyeyoung Ryu, Soyeon Kim, Dain Kim, Sooan Han, Keeheon Lee, and Younah Kang. 2020. Simple and steady interactions win the healthy mentality: designing a chatbot service for the elderly. *Proceedings of the ACM on human-computer interaction*, 4(CSCW2):1–25.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276, Online, November. Association for Computational Linguistics.

Ashish Sharma, Inna W. Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. 2021. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. In *Proceedings of the Web Conference 2021*, WWW '21, page 194–205, New York, NY, USA. Association for Computing Machinery.

Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. 2015. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28.

Aimee Spector, Lene Thorgrimsen, BOB Woods, Lindsay Royan, Steve Davies, Margaret Butterworth, and Martin Orrell. 2003. Efficacy of an evidence-based cognitive stimulation therapy programme for people with dementia: randomised controlled trial. *The British Journal of Psychiatry*, 183(3):248–254.


(c) 2024 中国中文信息学会计算语言学专业委员会

Seiki Tokunaga, Katie Seaborn, Kazuhiro Tamura, and Mihoko Otake-Matsuura. 2019. Cognitive training for older adults with a dialogue-based, robot-facilitated storytelling system. In Rogelio E. Cardona-Rivera, Anne Sullivan, and R. Michael Young, editors, *Interactive Storytelling*, pages 405–409, Cham. Springer International Publishing.

Seiki Tokunaga, Kazuhiro Tamura, and Mihoko Otake-Matsuura. 2021. A dialogue-based system with photo and storytelling for older adults: toward daily cognitive training. *Frontiers in Robotics and AI*, 8:644964.

Stefano Valtolina and Liliana Hu. 2021. Charlie: A chatbot to improve the elderly quality of life and to make them more active to fight their sense of loneliness. In *Proceedings of the 14th Biannual Conference of the Italian SIGCHI Chapter*, pages 1–5.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Yida Wang, Pei Ke, Yinhe Zheng, Kaili Huang, Yong Jiang, Xiaoyan Zhu, and Minlie Huang. 2020. A large-scale chinese short-text conversation dataset. In *Natural Language Processing and Chinese Computing: 9th CCF International Conference, NLPCC 2020, Zhengzhou, China, October 14–18, 2020, Proceedings, Part I 9*, pages 91–103. Springer.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

CCL 2024


(c) 2024 中国中文信息学会计算语言学专业委员会

# 基于思维链的跨语言多文档摘要生成技术研究

**祁天[1]，杨建安[2]，赵铁军[1]，杨沐昀[1]**
[1]哈尔滨工业大学 计算学部，中国 黑龙江 哈尔滨，150001
[2]东北大学 信息科学与工程学院，中国 辽宁 沈阳，110819
qitian@stu.hit.edu.cn，20206422@stu.neu.edu.cn，{tjzhao,yangmuyun}@hit.edu.cn

## 摘要

随着全球化的加速发展，跨语言信息的高效传递与理解变得尤为重要。传统的多文档摘要生成技术可以提升信息获取效率，然而往往忽视了跨语言场景下的特殊挑战。为了缓解这一问题，本文提出了跨语言多文档摘要生成任务。我们首先构建了一个全面的跨语言多文档摘要测试集作为评估基准，其次提出了一种基于思维链技术的跨语言多文档摘要生成方法，并对其进行了实验验证。在实验中，我们使用了几种典型的大语言模型，并通过人工评估和自动评估来验证我们的方法。结果表明，我们提出的基于思维链的方法在跨语言多文档摘要生成任务上取得了显著的性能提升，为解决语言障碍下的信息获取问题提供了有效的解决方案。

**关键词：** 多文档摘要；跨语言摘要；大语言模型；思维链

# Cross-lingual Multi-document Summarization Based on Chain-of-Thought

**Tian Qi[1], Jianan Yang[2], Tiejun Zhao[1], Muyun Yang[1]**
[1]Faculty of Computing, Harbin Institute of Technology, Harbin 150001,China
[2]College of Information Science and Engineering, Northeastern University, Shenyang,110819, China
qitian@stu.hit.edu.cn, 20206422@stu.neu.edu.cn, {tjzhao,yangmuyun}@hit.edu.cn

## Abstract

With the accelerating development of globalization, efficient transmission and understanding of cross-lingual information have become particularly important. Traditional multi-document summarization techniques can improve information retrieval efficiency but often overlook the challenges posed by cross-lingual scenarios. To address this issue, this paper proposes a task for cross-lingual multi-document summarization generation. We first construct a comprehensive cross-lingual multi-document summarization test set as an evaluation benchmark. Secondly, we propose a cross-lingual multi-document summarization generation method based on Chain-of-Thought technology and validate it through experiments. In the experiments, we use several typical large language models and verify our method through manual and automatic evaluation. The results show that our proposed method based on Chain-of-Thought significantly improves performance in cross-lingual multi-document summarization generation tasks, providing an effective solution for information retrieval issues under language barriers.

**Keywords:** Multi-Document Summarization , Cross-Lingual Summarization , Large Language Model , Chain-of-Thought

CCL 2024

## 1 引言

随着全球化的加速和信息技术的飞速发展，人们面临着信息过载的挑战。在这个背景下，多文档摘要技术成为了解决信息过载问题的关键工具之一。多文档摘要旨在从一组相关文档中提取和整合关键信息，生成简洁、连贯的摘要，帮助用户快速理解和获取大量文本资料的精髓。随着互联网上多语言内容的不断增长，跨语言多文档摘要（Cross-Lingual Multi-Document Summarization, CLMDS）的需求日益增加，它不仅要求模型能够处理多文档信息，还要求能够跨越语言障碍，从不同语言的文档中提取和整合信息。

传统的多文档摘要方法在处理单一语言的文档集合时已取得一定的进展，但在跨语言环境下，它们面临着新的挑战。不同语言之间的语法结构、词汇表达和文化背景差异使得跨语言信息提取和整合变得更加复杂。此外，不同语言之间的资源分布不均，如英语等资源丰富的语言拥有大量的训练数据，而其他语言则可能缺乏足够的数据支持，这进一步加剧了跨语言多文档摘要的难度。

近年来，大规模语言模型（Large language models, LLMs）如GPT-3、Llama等在自然语言处理（NLP）领域取得了显著的成就，它们在单文档摘要任务中表现出色，显示出强大的语言理解和生成能力。然而，当应用于多文档摘要，尤其是跨语言场景时，这些模型仍然面临着理解文档间复杂关系、生成高质量摘要的挑战。为了克服这些挑战，本文提出了基于思维链（Chain-of-Thought, CoT）的跨语言多文档摘要生成方法，该方法通过模拟人类解决问题时的逻辑推理过程，引导模型逐步生成摘要，从而提高了处理复杂任务的能力。本文通过人工构建数据集分析了该方法在处理多文档关系、提高摘要质量方面的潜力和挑战，并探讨未来的研究方向。本文主要有以下三个贡献:

- 提出了跨语言多文档摘要任务。明确了该任务的定义，为解决语言障碍下的信息获取问题提供了新的研究方向。
- 构建了跨语言多文档摘要测试集。我们通过人工的方式，构建了一个全面的跨语言多文档摘要测试集，为研究人员在该任务上提供了一个可靠的评估基准。
- 提出了基于思维链的跨语言多文档摘要生成方法，并通过实验验证了该方法在跨语言多文档摘要生成上的性能优势。

## 2 相关工作

### 2.1 多文档摘要

多文档摘要（Multi-Document Summarization, MDS）是指从同一主题下的多篇文档中提取出一份能够表达这些文档核心内容的综合性摘要(黄文彬and 倪少康, 2017)。与单文档摘要（Single Document Summarization, SDS）相比，多文档摘要提供了基于某个主题的文档集的重要内容概述，考虑了单个文档的特有信息及文档之间的冗余信息，极大地提高了主题相似的多篇文档的处理效率。因此，多文档摘要成为自动文本摘要领域的一个热点。Wang et al. (2020)提出了异构图神经网络，专门用于提取多文档摘要。这种异构图包括代表单词、句子和文档的节点。基于共现单词节点建立句子节点和文档节点之间的连接，促进了对多文档之间关系的更全面表示。Liu et al. (2018)利用Transformer进行多文档摘要任务，从一系列的参考文献中生成英文维基百科文章。在此工作的基础上，Liu and Lapata (2019)提出了一个基于Transformer结构以层级方式编码文档的模型，通过注意力机制表示跨文档之间的关系，允许多个文档信息共享而不是简单拼接文本为一个平行化的长序列。Li et al. (2020)提出一个利用图结构学习跨文档关系的端到端模型结构，该模型结构可以学习字级别的嵌入也可以学习到句子级别的嵌入。为了从不同粒度级别得到更丰富的语义表示，Jin et al. (2020)为抽取式和生成式多文档摘要提出一个多粒度交互网络模型，该模型能够联合学习词级别、句子级别和文档级别三个粒度的表示。然而目前多文档摘要的研究集中于单语言上，即输入文档集合与输出摘要是同一种语言，在跨语言上的研究较少 (Pontes et al., 2020)。

**基金项目**：国家自然科学基金项目“知识驱动的文本生成模型研究”(62376075)
©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

### 2.2 跨语言摘要

跨语言摘要（Cross-lingual summarization, CLS）是一项对源语言文档核心信息进行内容归纳，以目标语言的形式组织成摘要的任务。早期的工作通常集中在流水线方法上 (Leuski et al., 2003; Orăsan and Chiorean, 2008; Wan et al., 2010; Wan, 2011; Yao et al., 2015)，即先翻译再摘要或先摘要再翻译。随着深度学习技术在自然语言处理任务中的不断应用，许多研究者将研究重点转移到了端到端的跨语言摘要模型上。Zhu et al. (2019)使用round-trip翻译策略，构建了大规模的跨语言摘要平行语料库，并首次训练了基于神经网络的跨语言摘要模型。Zhu et al. (2020)提出了将Transformer模型与外部概率双语词典结合的融合翻译模式，降低了模型对于外部数据的过分依赖。Cao et al. (2020)使用一个多任务学习框架，共同学习摘要和上下文级别的对齐表示。Xu et al. (2020)采用多语言预训练的方法进行跨语言摘要任务，该方案利用大规模的预训练来提高跨语言摘要的质量。随着大模型研究的兴起，Wang et al. (2023b)探索了利用大模型引导生成跨语言摘要，通过使用各种提示引导大模型在端到端和流水线不同的范式上执行zero-shot生成跨语言摘要。然而，目前跨语言摘要的研究集中于单文档，针对多文档的跨语言摘要并没有系统性研究。

### 2.3 思维链

思维链（Chain-of-Thought，CoT）是将问题分解为一系列中间推理步骤的完整思考过程，形式上体现为输入→推理链→输出的映射 (Zhang et al., 2023)。这种方法通常比传统的直接推理更为有效。许多研究已经证明了思维链在广泛的领域中的有效性，包括算术推理、常识推理和符号推理 (Wei et al., 2022; Kojima et al., 2022; Wang et al., 2022)。Wang et al. (2023c)将思维链引入到摘要领域，提出了SumCoT，旨在引导大模型以逐步的方式生成摘要。首先指导大模型使用手动设计的问题提示从源文档中提取核心新闻要素。随后将提取的要素以及源文档中的额外细节整合起来，以产生全面且信息丰富的新闻摘要。

## 3 方法

### 3.1 问题描述

跨语言多文档摘要，是指从一组主题相近的多篇源语言文档中，生成以目标语言形式表达的、能够综合地概括这些文档核心内容的摘要。跨语言多文档摘要可以形式化描述为：给定一个多文档集合$D_{\text{src}} = \{d_{\text{src},1}, d_{\text{src},2}, ..., d_{\text{src},N}\}$，其中$N$是文档数量，$d_{\text{src},i} = \{s_{i,j}|j \in [1, M_{d_i}]\}$ 是用源语言编写的包含$M_{d_i}$个句子的文档，每个集合中的文档可能包含不同数量的句子。任务的目标是生成一个关于目标语言的多文档摘要$S_{\text{tgt}}$，使得该摘要可以用来概括整个源语言文档集合中的内容。

### 3.2 整体架构

图 1展示了本文提出的基于思维链的跨语言多文档摘要生成的整体架构。该方法包括三个主要阶段。首先，使用单文档摘要思维链（Chain1）对文档集合中的每一篇文档生成摘要；然后，通过多文档摘要思维链（Chain2）来组织多个文档之间的联系，生成关于源语言的多文档摘要；最后，通过跨语言思维链（Chain3）利用上述信息生成关于目标语言的跨语言多文档摘要。该方法可形式化为下面的公式，其中各参数的含义见后文：

$$CLMDS = \text{Chain3}(\text{Chain2}(\text{Chain1}(D; Q; A; P_1), P_2); L; P_3)$$

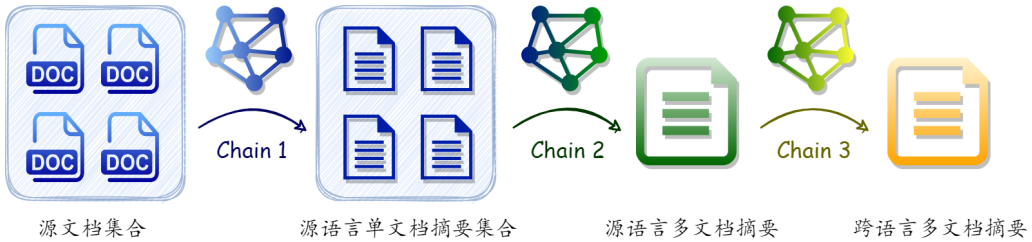



图 1. 基于思维链的跨语言多文档摘要生成整体架构


(c) 2024 中国中文信息学会计算语言学专业委员会

### 3.2.1 第一阶段：单文档摘要生成

基于思维链生成单文档摘要如图 2所示，将摘要任务拆分为以下两个主要步骤：

第一步是关键信息提取。针对学术论文的写作特点，我们创建了一系列引导性问题$Q = [q_1, q_2, q_3, q_4]$，以引导大模型提取论文中的核心内容，如研究对象、研究方法、研究结果等。假设源文档为$D_i$，输出的提取答案为$A_i$，则该步骤可形式化为：

$$A_i = LLM(D_i; Q)$$

第二步是多信息整合与摘要。在该步我们将提取的内容与源文档中更详细的信息进行整合，具体而言，将源文档、提取答案和整合生成摘要提示词$P_1$连接起来，以引导大模型生成该文档的摘要$SDS_i$，则该步骤可形式化为：

$$SDS_i = LLM(D_i; A_i; P_1)$$

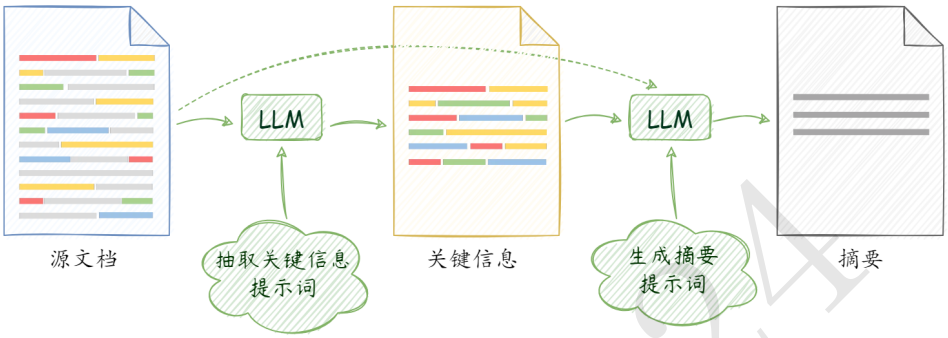



图 2. 单文档摘要生成思维链

### 3.2.2 第二阶段：多文档摘要生成

在本阶段，我们利用在第一阶段对每篇文档单独生成的摘要，通过思维链将这些摘要合并成一个整体的多文档摘要。合并过程可以采用不同的策略，例如简单直接地连接单篇文档的摘要后输入大模型；或者通过分层合并(Hierarchical merging)和增量更新(Incremental updating)的思维链策略进行生成。这两种策略已在书籍摘要领域进行了研究(Wu et al., 2021; Chang et al., 2023)，但目前尚未在多文档摘要领域进行深入探讨。

在层次化生成多文档摘要时，我们首先合并相邻两篇文档的摘要，然后逐层地合并摘要，直到只剩下一个摘要（整个文档集合的摘要），如图 3所示。这种方法类似于一种递归的过程，每一步都在合并低一层摘要以覆盖更多信息。层次化生成多文档摘要需要两个提示词引导大模型生成，一个用来合并单篇文档的摘要，另一个用来合并两个多文档集合子集的摘要。这种方法类似于将一个较大的任务分解成更小的部分，然后逐步合并这些部分来构建最终结果。这种策略的优点在于降低了模型处理的复杂性，与直接生成的方法相比，大幅降低了输入的上下文长度。我们将每篇文档的摘要作为叶子节点，构建初始的摘要节点集合$L = \{SDS_1, SDS_2, \ldots, SDS_n\}$，$P_2$为该阶段提示词，设BuildTree是一个递归函数，用于构建二叉树，则层次化生成多文档摘要的形式化公式可以表示为：

$$MDS = \text{BuildTree}(L, P_2)$$

在增量化生成多文档摘要时，我们首先生成一个全局摘要，然后逐步将每个单篇文档的摘要集成到这个全局摘要中，如图 4所示。这种方法类似于一种迭代的过程，每一步都在原有的摘要基础上进行更新，以适应新的信息。增量化生成多文档摘要需要两个提示词引导大模型生成，一个用来对初始的两篇文档生成一个全局摘要，另一个用来将新输入的单篇摘要合并到全局摘要中去。这种策略的优点在于可以更好地处理文档块之间的依赖关系，避免了一次性处理整个文档可能引入的不连贯性和错误。假设某文档集合中包含$n$篇文档，其中第$i$篇文档摘要为$SDS_i$，$P_2$为该阶段提示词，则增量化生成策略可形式化为：

$$MDS_0 = \emptyset, \forall i \in \{1, \ldots, n\}, MDS_i = \text{LLM}(MDS_{i-1}, SDS_i, P_2)$$


(c) 2024 中国中文信息学会计算语言学专业委员会

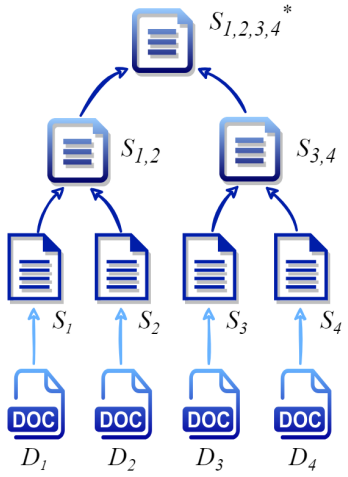



图 3. 层次化生成多文档摘要示意图

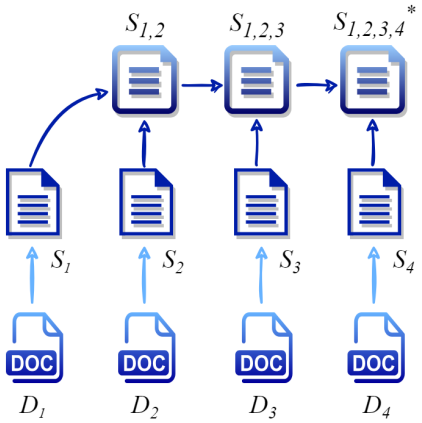



图 4. 增量化生成多文档摘要示意图

#### 3.2.3 第三阶段：跨语言多文档摘要生成

在这一阶段，我们利用第二阶段生成的关于源语言的多文档摘要，通过跨语言多文档摘要思维链生成关于目标语言的跨语言多文档摘要。具体来说，我们将源语言的多文档摘要$MDS$、文档集合中每篇文章的摘要$SDS_i$与跨语言摘要生成的提示词$P_3$连接起来，以引导大模型生成该文档集合的跨语言摘要$CLMDS$，即：

$$CLMDS = LLM(MDS; L; P_3)$$

其中$L = \{SDS_1, SDS_2, \ldots, SDS_n\}$，$n$为文档集合长度。

## 4 实验

### 4.1 数据集

针对本文提出的跨语言多文档摘要任务，目前并没有现成的相关数据集可供使用。因此，本文精心构建了一个新的测试集，旨在评估我们提出的方法在这一任务上的表现。我们选择了医学和人工智能两个特定领域的学术论文作为文档来源。为了保证数据集的质量和代表性，我们通过关键词筛选、时间范围限定和质量评估从Pubmed医学论文数据库[-1]和AAAI人工智能论文集[0]中分别选择了一定数量的医学和人工智能相关论文。同时，在创建多文档摘要测试集时，我们通过人工筛选的方式确保簇内文章主题保持一致。考虑到输入上下文长度的限制，我们对于每篇学术论文只截取了部分关键内容作为源文档输入。我们将此数据集公开至github供相关研究使用[1]。

我们构建的测试集包括单文档和多文档、单语言和跨语言的摘要任务。在构建过程中，我们邀请了三位领域专家根据完整的写作协议编写了源文档的专业摘要。编写摘要时，特别强调了摘要的全面性、客观性和风格的一致性，以确保其质量和可靠性。测试集的相关统计指标如表 1所示，#Sentence、#Words、#Tokens分别表示句子、单词/中文字符和tokens的平均长度，DOC表示源文档或源文档集合，Sum-en表示英文参考摘要，Sum-zh表示中文参考摘要。

### 4.2 实验设置

为了验证本文提出的方法在跨语言多文档摘要任务上的有效性，我们选择了几个具有代表性的大语言模型进行实验：

- ChatGLM (Du et al., 2022)是智谱AI研发的一个开源的、支持中英双语的对话语言模型，基于General Language Model (GLM) 架构，具有6.2B参数。在本文中我们使用的是chatglm3-6b-32k模型 [2]。

[-1]https://pubmed.ncbi.nlm.nih.gov/
[0]https://ojs.aaai.org/index.php/AAAI/issue/archive
[1]https://github.com/yja2576/CLMDS-Dataset
[2]https://huggingface.co/THUDM/chatglm3-6b-32k


(c) 2024 中国中文信息学会计算语言学专业委员会

| Task | Domain | #Docs | #Cluster | #Sentences | | | #Words | | | #Tokens | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Doc | Sum-en | Sum-zh | Doc | Sum-en | Sum-zh | Doc | Sum-en | Sum-zh |
| SDS | AI | 1.00 | 60 | 51.23 | 7.48 | 4.90 | 1,103.93 | 196.83 | 246.28 | 1,335.82 | 223.05 | 253.77 |
| | Medical | 1.00 | 61 | 18.33 | 4.15 | 4.23 | 474.69 | 161.29 | 320.57 | 630.52 | 197.39 | 375.31 |
| | Overall | 1.00 | 121 | 34.64 | 5.80 | 4.56 | 786.71 | 178.91 | 283.27 | 980.26 | 210.11 | 315.04 |
| MDS | AI | 2.52 | 50 | 125.36 | 7.34 | 7.34 | 2,712.72 | 261.96 | 496.78 | 3,274.94 | 296.22 | 514.42 |
| | Medical | 2.72 | 50 | 49.48 | 6.82 | 6.74 | 1,302.00 | 273.28 | 514.16 | 1,749.58 | 328.82 | 628.28 |
| | Overall | 2.62 | 100 | 87.42 | 7.08 | 7.04 | 2,007.36 | 257.62 | 505.47 | 2,512.26 | 312.52 | 584.80 |

Table 1: 本文构建的跨语言多文档摘要任务测试集的数据统计情况

- Qwen（通义千问）(Bai et al., 2023)是阿里云发布的大语言模型，参数规模从1.8B到72B，在本文中我们使用的是Qwen1.5-7B-Chat [3]模型。
- Llama-2 (Touvron et al., 2023)是Meta推出的一种Decoder-Only架构的大模型，模型参数范围从7B到70B。在本文中我们使用Llama-2的7B模型 [4]，当涉及跨语言任务时，我们使用了基于Llama-2的中文微调模型 [5]。
- GPT-3.5-turbo是OpenAI发布的GPT-3.5系列模型 (Ouyang et al., 2022)之一。该模型可以理解和生成自然语言或代码，并经过优化以在聊天中使用，但也适用于非聊天任务。本文使用的版本是*gpt-3.5-turbo-0125*。

为了全面评估性能情况，我们选择了多种评价指标。首先，我们考虑了基于参考摘要的指标。其中，ROUGE (Lin, 2004)用于衡量生成摘要与参考摘要之间的词重叠程度，我们主要关注ROUGE-1、ROUGE-2和ROUGE-L指标。另外，我们还采用了BERTScore (Zhang et al., 2019)），它利用BERT (Devlin et al., 2018)嵌入来衡量语义相似性。此外，在无参考摘要的情况下，我们利用GPT-3.5（gpt-3.5-turbo-0125）作为摘要评估器。GPT-3.5被证明是一个强大的自然语言生成评估器 (Jain et al., 2023; Wang et al., 2023a; Shen et al., 2023)，我们向大模型提供源文本和生成的摘要（适用于GPT-3.5的16k上下文窗口），并要求输出一个从1到10的分数，然后乘以10得到最终的分数。完整的提示模板见附录A。

同时，本文进行了人工评价，综合考虑以往的工作所评价的维度，我们从以下四个维度对生成的摘要进行了评估：

1. 流畅性（Fluency）：评估摘要的自然度和可读性。一个流畅的摘要应该具有正确的句子结构、无语法错误、用词恰当，读起来通顺。
2. 连贯性（Coherence）：评估摘要中信息的逻辑顺序。一个连贯的摘要应该具有清晰的结构，句子之间有合理的过渡，没有逻辑上的跳跃或混乱。
3. 一致性（Consistency）：评估摘要中的事实、观点和信息是否与原文一致。摘要应该准确地反映原文的内容，不包含原文中未提及的信息，也不包含错误的信息。
4. 相关性（Relevance）：评估摘要是否聚焦原文最重要的信息，排除不相关或次要的细节。

### 4.3 实验结果与分析

#### 4.3.1 第一阶段：单文档摘要生成

我们通过Zero-shot和Chain-of-Thought两种生成方式生成单文档摘要，实验结果如表 2所示。在该表中，带有灰色背景的数据 表示Zero-shot和Chain-of-Thought两种生成方式中得分较高的一种。观察实验结果发现，在使用思维链后，摘要整体质量有了明显提升。在整个实验中，GPT-3.5模型在多个领域和指标下都表现出色，这可能归因于其在自然语言生成领域的强大性能。

[3]https://huggingface.co/Qwen/Qwen1.5-7B-Chat
[4]https://huggingface.co/meta-llama/Llama-2-7b-chat-hf
[5]https://huggingface.co/FlagAlpha/Llama2-Chinese-7b-Chat

具体而言，对于医学领域，所有模型在使用思维链生成时，各项指标均比Zero-shot方式有所提升。这表明，思维链在医学领域的单文档摘要生成中是有效的。对于人工智能领域，尽管在一些指标上（如Qwen、Llama和GPT-3.5上的ROUGE-1/2/L），使用思维链生成方式并未带来得分提升。然而，根据 Goyal et al. (2022)的研究发现，自动评估结果与人类偏好并不一致。实际上，尽管GPT-3.5生成的摘要在自动评估指标上表现不佳，但其更受人类喜爱。

| Dataset | Model | Zero-shot | | | | | Chain-of-Thought | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | R-1 | R-2 | R-L | BS | GPT-3.5 | R-1 | R-2 | R-L | BS | GPT-3.5 |
| Medical | ChatGLM3-6b | 40.54 | 18.65 | 25.68 | 88.60 | 80.49 | 46.86 | 23.18 | 32.78 | 90.62 | 85.08 |
| | QWen1.5-7b | 46.70 | 22.19 | 32.64 | 91.00 | 85.57 | 47.40 | 23.13 | 34.92 | 91.42 | 87.21 |
| | Llama2-7b | 47.06 | 23.70 | 33.00 | 90.84 | 86.39 | 47.53 | 24.18 | 33.09 | 90.93 | 86.72 |
| | GPT-3.5-turbo | **52.39** | **28.10** | **37.92** | **92.21** | **87.38** | **53.21** | **29.70** | **40.06** | **92.31** | **88.20** |
| AI | ChatGLM3-6b | 31.55 | 12.12 | 18.84 | 86.22 | 83.00 | **36.89** | **14.32** | **21.71** | 87.74 | 88.00 |
| | QWen1.5-7b | 32.30 | 8.81 | 18.02 | 87.52 | 88.50 | 30.29 | 7.82 | 16.69 | 87.47 | 89.33 |
| | Llama2-7b | **35.97** | **13.53** | **21.71** | 87.58 | 87.83 | 35.58 | 13.45 | 20.97 | 87.78 | 88.17 |
| | GPT-3.5-turbo | 33.84 | 10.58 | 19.50 | **87.91** | **89.17** | 33.62 | 11.45 | 19.58 | **87.79** | **89.67** |

Table 2: 不同方法在单文档摘要测试集上的实验结果

因此，为验证生成摘要的质量，我们进行了人工评估。我们从流畅性、连贯性、一致性和相关性四个维度（具体标准参见第 4.2节）对摘要进行人工打分，评分结果见图 5和图 6。观察图表结果，我们发现不同大模型在使用了思维链技术后，生成的摘要在四个维度上都有不同程度的分数提升。其中，流畅性和一致性的提升较小，而连贯性和相关性的提升较为显著。这主要归因于我们设计了一个逐步引导大模型的提示词(Prompt)，使其遵循人类生成摘要的逻辑来构建摘要，并指示其需要保留哪些重要内容（例如，研究的对象、背景、方法、结果等）。这一做法在一定程度上弥补了大模型在逻辑组织和内容重要度判断能力上的不足。人工评估再次证明了思维链技术（Chain-of-Thought）的有效性。

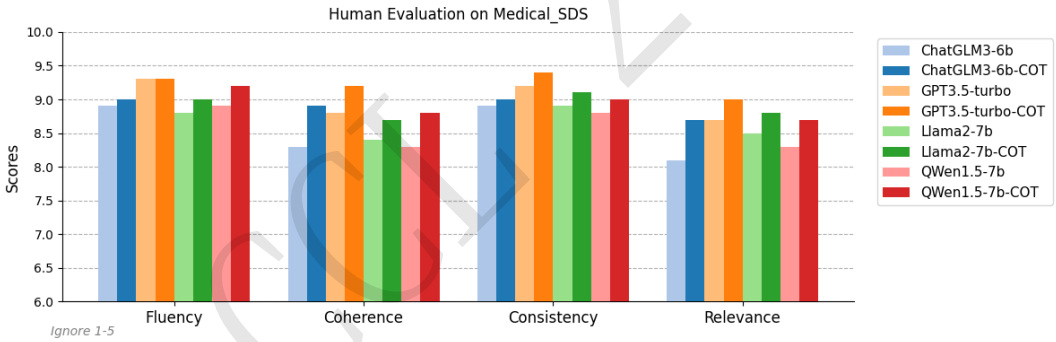



图 5. 医学领域单文档摘要生成结果的人工评分

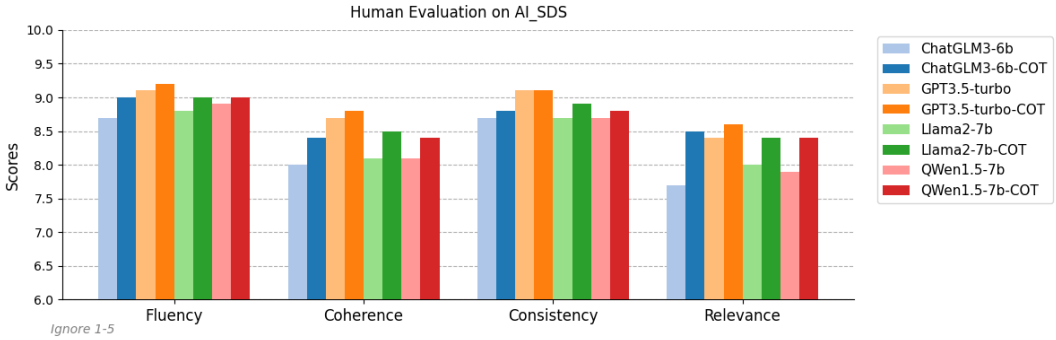



图 6. 人工智能领域单文档摘要生成结果的人工评分

#### 4.3.2 第二阶段：多文档摘要生成

在第二阶段的多文档摘要生成任务中，我们采用了三种不同的生成方法：直接生成（Direct）、层次化生成（Hierarchical）和增量化生成（Incremental）。这些方法利用第一阶段不同方式生成的单文档摘要，其中w/o表示第一阶段Zero-shot生成的单文档摘要，w/表示第一阶段Chain-of-Thought生成的单文档摘要，并据此生成多文档摘要。表 3展示了在医学和人工智能多文档摘要测试集（Medical MDS和AI MDS）上使用以上三种方法进行实验的结果。带有灰色背景的数据表示在每一种生成方法下，w/o与w/生成分数较高的结果。

| Metric | Model | Medical MDS | | | | | | AI MDS | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Direct | | Hierarchical | | Incremental | | Direct | | Hierarchical | | Incremental | |
| | | w/o | w/ | w/o | w/ | w/o | w/ | w/o | w/ | w/o | w/ | w/o | w/ |
| ROUGE-1 | ChatGLM3-6b | 31.90 | 34.03 | 33.64 | 34.44 | 33.19 | **35.41** | 29.56 | 30.51 | 29.58 | 31.06 | 29.82 | **31.96** |
| | QWen1.5-7b | 38.45 | 38.21 | 39.07 | 38.08 | 38.93 | **39.48** | 33.77 | 35.14 | 33.75 | **35.46** | 33.97 | 34.93 |
| | Llama2-7b | 38.87 | 38.93 | 38.51 | 38.33 | **39.70** | 38.95 | 33.41 | **34.92** | 32.99 | 34.34 | 34.22 | 34.80 |
| | GPT-3.5-turbo | 40.10 | 39.77 | 41.08 | 40.00 | **42.80** | 42.33 | 34.73 | 34.47 | 33.89 | 34.28 | **35.32** | 35.26 |
| ROUGE-2 | ChatGLM3-6b | 11.38 | 13.21 | 12.27 | 13.45 | 12.16 | **13.89** | 8.22 | 9.63 | 8.38 | **10.11** | 8.20 | 10.10 |
| | QWen1.5-7b | 14.80 | 14.64 | 15.44 | 14.50 | 15.06 | **15.46** | 10.16 | 10.53 | 10.21 | **10.87** | 10.02 | 10.53 |
| | Llama2-7b | **16.65** | 16.58 | 16.36 | 16.22 | 16.62 | 16.57 | 11.35 | **11.79** | 11.20 | 11.35 | 11.64 | 11.52 |
| | GPT-3.5-turbo | 15.66 | 15.13 | 16.69 | 15.99 | **18.25** | 17.95 | 11.46 | 11.37 | 11.32 | 11.01 | 11.55 | **11.73** |
| ROUGE-L | ChatGLM3-6b | 18.97 | 19.98 | 19.53 | **20.57** | 18.98 | 20.16 | 17.45 | 18.44 | 17.57 | **18.70** | 17.30 | 18.41 |
| | QWen1.5-7b | 23.41 | 23.01 | 23.64 | 23.15 | 23.19 | **23.67** | 19.48 | 20.05 | 19.91 | **20.52** | 18.59 | 19.37 |
| | Llama2-7b | 24.12 | 23.60 | **24.17** | 23.55 | 24.01 | 23.49 | 20.83 | **21.52** | 20.66 | 21.29 | 20.16 | 20.62 |
| | GPT-3.5-turbo | 24.41 | 24.03 | 25.27 | 24.14 | **25.99** | 25.83 | **21.25** | 20.74 | 19.90 | 20.39 | 20.22 | 20.39 |
| GPT-3.5 | ChatGLM3-6b | 76.80 | **81.60** | 77.40 | 80.00 | 78.20 | 81.20 | 82.80 | 86.00 | 82.40 | **86.60** | 83.20 | 84.80 |
| | QWen1.5-7b | 86.20 | 85.60 | **86.80** | 86.00 | 84.60 | 84.20 | 86.80 | 87.60 | 87.80 | **88.20** | 85.40 | 86.40 |
| | Llama2-7b | 83.80 | **86.00** | 84.40 | 84.60 | 85.00 | 85.40 | 87.20 | 87.40 | 87.00 | **87.60** | 86.40 | 87.20 |
| | GPT-3.5-turbo | 86.00 | 86.60 | 87.00 | **87.20** | 86.20 | 86.20 | 88.60 | 88.80 | 88.20 | **89.20** | 88.40 | 89.00 |

Table 3: 不同方法在多文档摘要测试集上的实验结果

我们对表 3中使用第一阶段生成结果得到的三种策略下，不同模型生成的多文档摘要指标进行了平均值计算和分析，结果如表 4所示。针对医学领域的多文档摘要生成任务，层次化生成策略相较于直接生成策略在ROUGE-2和ROUGE-L指标上均有所提升，仅在ROUGE-1略微下降，而增量化生成策略的提升效果更为显著。对于人工智能领域，层次化生成策略在所有指标上均表现出提升，而增量化生成策略在ROUGE-L指标上的降低可能与生成文本长度减少有关。总体而言，层次化和增量化生成策略均表现出更好的性能，显著提高了多文档摘要生成的质量，这进一步证明了思维链方法在多文档摘要生成中的有效性。

| Metric | Medical MDS | | | AI MDS | | |
|---|---|---|---|---|---|---|
| | Direct | Hierarchical | Incremental | Direct | Hierarchical | Incremental |
| ROUGE-1 | 37.74 | 37.71(↓0.03) | **39.04(↑1.30)** | 33.76 | **33.79(↑0.02)** | **34.24(↑0.48)** |
| ROUGE-2 | 14.89 | **15.04(↑0.15)** | **15.97(↑1.08)** | 10.83 | **10.84(↑0.01)** | **10.97(↑0.14)** |
| ROUGE-L | 22.66 | **22.85(↑0.19)** | **23.29(↑0.63)** | 20.19 | **20.23(↑0.04)** | 19.70(↓0.49) |

Table 4: 不同生成策略在第二阶段的多文档摘要生成结果比较

| Metric | Medical MDS | | | AI MDS | | |
|---|---|---|---|---|---|---|
| | Direct | Hierarchical | Incremental | Direct | Hierarchical | Incremental |
| ROUGE-1 | **+0.4050** | -0.3625 | **+0.3875** | **+0.8925** | **+1.2325** | **+0.9050** |
| ROUGE-2 | **+0.2675** | -0.1500 | **+0.4450** | **+0.5325** | **+0.5575** | **+0.6175** |
| ROUGE-L | -0.0725 | -0.3000 | **+0.2450** | **+0.4350** | **+0.7150** | **+0.6300** |

Table 5: 第一阶段思维链在第二阶段生成中使用情况的效果分析


(c) 2024 中国中文信息学会计算语言学专业委员会

从表 3的结果还可以清晰看出，使用第一阶段生成的思维链摘要（w/）后，大多数模型在所有生成策略下的性能都有所提升，尤其在人工智能领域的测试集中表现更加显著。进一步的计算发现，如表 5所示，在第二阶段使用第一阶段思维链后生成结果的分数变化。相对于直接生成单文档摘要的方法（w/o），使用第一阶段思维链生成的摘要（w/）在人工智能领域的平均提升分数更高，这一趋势在医学领域的直接生成策略和增量化方法中也得到了体现。

为了深入研究在医学多文档摘要测试集上，第一阶段思维链在层次化生成策略下性能降低的原因，我们进行了人工评估，评估结果如表 6所示，w/o表示没有使用第一阶段的思维链。结果显示，第一阶段生成的思维链摘要在第二阶段的多文档摘要生成任务中性能显著提升，表明思维链对于帮助模型理解文本内容并生成更准确、连贯的摘要起到了关键作用。人工评价再一次验证了 Goyal et al. (2022)的研究结论，即自动评估结果与人类偏好的不一致性。

| Model | Medical_MDS | AI_MDS |
|---|---|---|
| | *Flu/Coh/Con/Rel* | *Flu/Coh/Con/Rel* |
| ChatGLM3-6b | **9.10/8.70/8.60/8.40** | **8.80/8.50/8.70/8.40** |
| w/o chain1 | 8.90/8.70/8.10/8.00 | 8.70/8.40/8.50/7.80 |
| QWen1.5-7b | **8.80/8.70/8.80/8.50** | **9.00/8.80/8.90/8.50** |
| w/o chain1 | 8.80/8.50/8.30/8.20 | 8.70/8.60/8.70/8.10 |
| Llama2-7b | **9.00/8.90/8.70/8.60** | **8.90/9.00/8.80/8.70** |
| w/o chain1 | 8.80/8.60/8.40/8.10 | 8.80/8.70/8.70/8.30 |
| GPT-3.5-turbo | **9.00/8.90/8.90/8.70** | **9.20/9.00/9.10/8.80** |
| w/o chain1 | 9.00/8.90/8.80/8.60 | 9.10/9.00/8.80/8.80 |

Table 6: 在医学和人工智能领域多文档摘要测试集上的层次化生成策略的人工评价结果

### 4.3.3 第三阶段：跨语言多文档摘要生成

在本节中，我们利用 4.1节构建的跨语言多文档摘要的测试集（CLMDS）进行了跨语言多文档摘要生成的实验。我们分别评估了两个模型（ChatGLM3-6b和GPT-3.5-turbo）在不同条件下的性能，包括是否使用了三个阶段的思维链生成。实验结果如表 7 所示。

我们对未使用思维链直接生成的方法，以及使用了第一阶段和第三阶段思维链的方法，以及使用了三个阶段所有思维链的方法进行了比较。结果显示，在医学领域的CLMDS任务中，无论是在ROUGE-1、ROUGE-2、ROUGE-L还是GPT-3.5得分方面，使用了思维链的模型都表现出了显著的提升。特别是当使用了所有三个阶段的思维链时，性能提升更为明显，这进一步验证了思维链在帮助模型理解文本内容并生成更准确、连贯的摘要方面的有效性。对于人工智能领域的CLMDS任务，我们观察到类似的趋势。使用思维链的模型在所有指标下均实现了提升，尤其是在使用了所有三个阶段思维链的情况下，性能得到了最大程度的提升。这些结果强调了思维链方法在跨语言多文档摘要生成任务中的有效性和重要性。

综上所述，本阶段的实验结果表明，在跨语言多文档摘要生成任务中，思维链技术对于提升模型性能具有积极作用，可以有效改善生成摘要的质量。

| Model | Medical CLMDS | | | | AI CLMDS | | | |
|---|---|---|---|---|---|---|---|---|
| | R-1 | R-2 | R-L | GPT-3.5 | R-1 | R-2 | R-L | GPT-3.5 |
| ChatGLM3-6b | 35.82 | 11.94 | 22.13 | 68.60 | 36.48 | 11.68 | 22.29 | 79.80 |
| w/ chain1,3 | 37.79 | 13.05 | 22.79 | 69.40 | 37.36 | 12.30 | 22.75 | 80.00 |
| w/ chain1,2,3 | **46.47** | **18.40** | **27.67** | **73.60** | **44.68** | **15.82** | **25.31** | **80.60** |
| GPT-3.5-turbo | 22.43 | 9.08 | 14.89 | 76.80 | 38.26 | 14.06 | 26.14 | 80.40 |
| w/ chain1,3 | 40.68 | 17.22 | 26.43 | 78.80 | 42.41 | 16.46 | 27.69 | 80.60 |
| w/ chain1,2,3 | **49.11** | **21.97** | **29.94** | **81.60** | **45.48** | **18.24** | **27.69** | **82.20** |

Table 7: 使用不同思维链在跨语言多文档摘要生成的实验结果


(c) 2024 中国中文信息学会计算语言学专业委员会

## 5 结论

在本文中，我们提出了一种基于思维链的跨语言多文档摘要方法，并在医学和人工智能领域的学术论文上进行了实验验证。我们构建了一个新的测试集，用于评估我们方法在这一任务上的表现，并选择了一些典型的大语言模型进行实验。实验结果表明，无论是在单文档摘要还是多文档摘要生成阶段，无论是单语言还是跨语言，思维链技术都能显著提高摘要的质量。

这些发现对于跨语言多文档摘要领域具有重要意义。首先，我们的方法为解决跨语言多文档摘要生成任务提供了一种有效的思路。其次，我们构建的跨语言多文档摘要测试集也可为未来的研究提供参考。此外，我们的实验结果表明，不同的大语言模型在跨语言多文档摘要任务上表现不同，因此未来可以进一步研究模型选择和优化。

在未来的工作中，我们将继续优化思维链技术，探索更好的生成方法和模型选择策略。我们还计划扩展测试集，涵盖更多领域和语言，以更全面地评估方法的通用性和鲁棒性。我们相信，这项工作将为跨语言多文档摘要领域的研究和应用提供有价值的参考和启示。

## 参考文献

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Yue Cao, Hui Liu, and Xiaojun Wan. 2020. Jointly learning to align and summarize for neural cross-lingual summarization. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 6220–6231.

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2023. Booookscore: A systematic exploration of book-length summarization in the era of llms. *arXiv preprint arXiv:2310.00785*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356*.

Sameer Jain, Vaishakh Keshava, Swarnashree Mysore Sathyendra, Patrick Fernandes, Pengfei Liu, Graham Neubig, and Chunting Zhou. 2023. Multi-dimensional evaluation of text summarization with in-context learning. *arXiv preprint arXiv:2306.01200*.

Hanqi Jin, Tianming Wang, and Xiaojun Wan. 2020. Multi-granularity interaction network for extractive and abstractive multi-document summarization. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 6244–6254.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Anton Leuski, Chin-Yew Lin, Liang Zhou, Ulrich Germann, Franz Josef Och, and Eduard Hovy. 2003. Cross-lingual c* st* rd: English access to hindi information. *ACM Transactions on Asian Language Information Processing (TALIP)*, 2(3):245–269.

Wei Li, Xinyan Xiao, Jiachen Liu, Hua Wu, Haifeng Wang, and Junping Du. 2020. Leveraging graph to improve abstractive multi-document summarization. *arXiv preprint arXiv:2005.10043*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.


(c) 2024 中国中文信息学会计算语言学专业委员会

Yang Liu and Mirella Lapata. 2019. Hierarchical transformers for multi-document summarization. *arXiv preprint arXiv:1905.13164*.

Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. *arXiv preprint arXiv:1801.10198*.

Constantin Orăsan and Oana Andreea Chiorean. 2008. Evaluation of a cross-lingual romanian-english multi-document summariser.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Elvys Linhares Pontes, Stéphane Huet, Juan-Manuel Torres-Moreno, and Andréa Carneiro Linhares. 2020. Compressive approaches for cross-language multi-document summarization. *Data & Knowledge Engineering*, 125:101763.

Chenhui Shen, Liying Cheng, Yang You, and Lidong Bing. 2023. Are large language models good evaluators for abstractive summarization? *arXiv preprint arXiv:2305.13091*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Xiaojun Wan, Huiying Li, and Jianguo Xiao. 2010. Cross-language document summarization based on machine translation quality prediction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 917–926.

Xiaojun Wan. 2011. Using bilingual information for cross-language document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1546–1555.

Danqing Wang, Pengfei Liu, Yining Zheng, Xipeng Qiu, and Xuanjing Huang. 2020. Heterogeneous graph neural networks for extractive document summarization. *arXiv preprint arXiv:2004.12393*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023a. Is chatgpt a good nlg evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048*.

Jiaan Wang, Yunlong Liang, Fandong Meng, Beiqi Zou, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2023b. Zero-shot cross-lingual summarization via large language models. *arXiv preprint arXiv:2302.14229*.

Yiming Wang, Zhuosheng Zhang, and Rui Wang. 2023c. Element-aware summarization with large language models: Expert-aligned evaluation and chain-of-thought method. *arXiv preprint arXiv:2305.13412*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Jeff Wu, Long Ouyang, Daniel M Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. 2021. Recursively summarizing books with human feedback. *arXiv preprint arXiv:2109.10862*.

Ruochen Xu, Chenguang Zhu, Yu Shi, Michael Zeng, and Xuedong Huang. 2020. Mixed-lingual pretraining for cross-lingual summarization. *arXiv preprint arXiv:2010.08892*.

Jin-ge Yao, Xiaojun Wan, and Jianguo Xiao. 2015. Phrase-based compressive cross-language summarization. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 118–127.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

CCL 2024

Zhuosheng Zhang, Yao Yao, Aston Zhang, Xiangru Tang, Xinbei Ma, Zhiwei He, Yiming Wang, Mark Gerstein, Rui Wang, Gongshen Liu, et al. 2023. Igniting language intelligence: The hitchhiker's guide from chain-of-thought reasoning to language agents. *arXiv preprint arXiv:2311.11797*.

Junnan Zhu, Qian Wang, Yining Wang, Yu Zhou, Jiajun Zhang, Shaonan Wang, and Chengqing Zong. 2019. Ncls: Neural cross-lingual summarization. *arXiv preprint arXiv:1909.00156*.

Junnan Zhu, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2020. Attend, translate and summarize: An efficient method for neural cross-lingual summarization. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics*, pages 1309–1321.

黄文彬and 倪少康. 2017. 多文档自动摘要方法的进展研究. 情报科学, 35:160–165.

# 附录

## A 本文中用到的提示词

### A.1 第一阶段：单文档摘要生成

**1.抽取关键信息提示词:**

*Read and analyze the preceding content carefully, anwser the question below:*

*what is the subject of the study?*
*what are the research methods of the study?*
*what are the primary findings of the study?*
*what are the innovations introduced in the study?*

**2.生成摘要提示词:**

*Let's integrate the above information and summarize the study in just one paragraph:*

### A.2 第二阶段：多文档摘要生成

**1.直接生成提示词:**

*Summary 1:*
{*summary_1*}

*Summary 2:*
{*summary_2*}

*Summary n:*
{*summary_n*}

*Please carefully review the provided summaries. Analyze the similarities and differences between these studies. Subsequently, creating a cohesive, unified summary in just one paragraph. This new summary should be crafted in the style of a survey. Remember, when referencing the studies, avoid using labels like "Summary n." Instead, directly address the content of each research study.*

*New summary:*

**2.层次化生成提示词:**

*Consider two sets of scientific articles. Your task is to merge the summaries of these articles into a single comprehensive summary, capturing the connections between different documents and presenting the merged summary in one paragraph.*

*Summary of the first set of articles:*

{*summary_1*}

CCL 2024

*Summary of the second set of articles:*

{*summary_2*}

*Merging Process:*
*1.Capturing Connections: Identify and analyze the connections between different documents in both summaries. Look for common themes, findings, methodologies, or implications.*
*2.Creating a One-Paragraph Summary: Merge the summaries of both sets of articles into a single paragraph that flows logically and smoothly. Ensure that the merged summary effectively communicates the overarching insights and contributions of the entire document collection.*

*Objective: Your goal is to create a merged multi-document summary in one paragraph that effectively captures the connections between the articles in both sets and presents them in a coherent and cohesive manner.*

*Updated summary:*

**3.增量化生成提示词:**
*Below is a summary of a science article:*

{*science_article_summary*}

*The following is a summary of the many scientific articles that have been generated:*

{*overall_summary*}

*Your task is to merge the summary of the science article into the overall multi-document summary. Ensure that the merge is done in a way that maintains coherence and consistency with the overall summary. Introduce any new concepts, terms, or references if they are being mentioned for the first time in the summary.*

*Step-by-step Process:*
*1.Merge: Incorporate the summary of the science article into the overall multi-document summary as a single paragraph. 2.Similarities and Differences: Analyze the similarities and differences between the articles, if applicable, and integrate these observations into the multi-document summary to create a smooth and logical narrative.*

*Objective: Your goal is to create a comprehensive, cohesive, and informative multi-document summary presented as a single paragraph that accurately represents the key aspects of the science article while maintaining coherence with the overall summary.*

*Updated summary:*

## A.3 第三阶段：跨语言多文档摘要生成

*Summary 1:*
{*summary_1*}

*Summary 2:*
{*summary_2*}

*Summary n:*
{*summary_n*}

CCL 2024

*Multi-document Summary:*
*{Multi-document Summary}*

*Integrate the summaries of each document and their multi-document summaries together, and generate summaries of these documents in Chinese. During the integration process, identify common themes, findings, methodologies, or impacts to ensure coherence and consistency. Introduce new concepts, terms, or references in the summary if they are mentioned for the first time.*

*Objective: Your goal is to create a comprehensive, coherent, and informative multi-document summary in Chinese that accurately reflects the main content of the articles mentioned above.*

*Updated summary:*

### A.4 单文档摘要评估提示词

*Score the following summary generated by another system given the source on a scale from 1 to 10 with regards to overall general summary quality. 1-point indicates a low quality summary, and 10-points indicates a very high quality summary. A high quality summary is grammatical, fluent, informative, relevant, coherent and factually consistent with the source.*

*Let's think step-by-step and just output one number representing the score.*

*Source:*
*{source_text}*

*Summary:*
*{summary_llm}*

*Your score: (What you must observe is that your output must contain only a number and no other content)*

### A.5 多文档摘要评估提示词

*Score the following multi-document summary generated by another system given the source on a scale from 1 to 10 with regards to overall general summary quality. 1-point indicates a low-quality multi-document summary, and 10-points indicates a very high-quality multi-document summary. A high-quality multi-document summary is concise in language, grammatical, fluent, informative, relevant, logically coherent and factually consistent with the source, successfully integrates all documents, and contains important content from all documents.*

*Furthermore, when you are scoring the summary, be strict and demand high standards, don't be too conservative with your scoring.*

*Let's think step-by-step and just output one number representing the score.*

*Source:*
*{source_text}*

*Multi-Document Summary:*
*{summary_llm}*

*Your score: (What you must observe is that your output must contain only a number and no other content)*

### A.6 跨语言多文档摘要评估提示词

*Score the following Chinese multi-document summary generated by another system given the English source on a scale from 1 to 10 with regards to overall general summary quality.*

CCL 2024

*1-point indicates a low-quality multi-document summary, and 10-points indicates a very high-quality multi-document summary. A high-quality multi-document summary is concise in language, grammatical, fluent, informative, relevant, logically coherent and factually consistent with the source, successfully integrates all documents, and contains important content from all documents.*

*Furthermore, when you are scoring the summary, be strict and demand high standards, don't be too conservative with your scoring.*

*Let's think step-by-step and just output one number representing the score.*

*English Source:*
*{source_text}*

*Chinese Multi-Document Summary:*
*{summary_llm}*

*Your score: (What you must observe is that your output must contain only a number and no other content)*

## B 生成实例分析

**基于思维链的单文档摘要生成示例1：**

**Source Document:**

*Title: Ranking age-specific modifiable risk factors for cardiovascular disease and mortality: evidence from a population-based longitudinal study*

*Cardiovascular disease (CVD) remains a paramount contemporary health challenge.....*

**ChatGLM3 Without COT:**

*Cardiovascular disease (CVD) remains a major health challenge, and this study examined age-specific effects of 14 risk factors on CVD and mortality in different age groups. The study analyzed data from 226,759 CVD-free participants aged 40 years and older in the UK Biobank during the period from baseline time (2006-2010) to September 30, 2021. The primary CVD outcome was a composite of incident coronary artery disease, heart failure, and stroke. The study found that age-specific disparities were observed in the risk factors contributing to CVD, and the overall PAF declined with age. Metabolic factors had the highest PAF in each age group, with hypertension (14.04% of the PAF) and abdominal obesity (9.58% of the PAF) being prominent. Behavioral factors had the highest PAF in the middle-aged group (10.68% of the PAF), and smoking was the leading behavioral factor in all age groups. In socioeconomic and psychosocial risk clusters, low income contributed most among middle-aged (3.74% of the PAF) and elderly groups (3.66% of the PAF), while less education accounted more PAF for quinquagenarian group (4.46% of the PAF). The study concluded that a large fraction of CVD cases and deaths were associated with modifiable risk factors in all age groups, and targeted efforts should focus on the most impactful risk factors, as well as age-specific modifiable risk factors. These findings may inform the development of more precise medical strategies to prevent and manage CVD and related mortality.*

**ChatGLM3 With COT:**

*The study examined the age-specific effects of 14 modifiable risk factors on cardiovascular disease (CVD) and mortality in a large sample of 226,759 CVD-free participants aged 40 years and older in the UK Biobank. The study found that a large fraction of CVD cases and deaths were associated with modifiable risk factors in all age groups. Metabolic factors had the highest PAF in each age group, with hypertension and abdominal obesity being prominent. Behavioral factors had the highest PAF in the middle-aged group, and smoking was the leading behavioral factor in*

CCL 2024

*all age groups. In socioeconomic and psychosocial risk clusters, low income and less education contributed most among middle-aged and elderly groups, respectively. The study highlights the need for targeted efforts to focus on the most impactful risk factors and age-specific modifiable risk factors to inform the development of more precise medical strategies to prevent and manage CVD and related mortality.*

**Reference:**

*This study investigates the age-specific impacts of 14 modifiable risk factors on cardiovascular disease (CVD) and mortality by analyzing data from 226,759 CVD-free participants aged 40 and above from the UK Biobank, spanning from 2006-2010 to September 2021. It identifies significant age-related disparities in the contribution of these risk factors to CVD, revealing that metabolic factors, especially hypertension and abdominal obesity, are the most influential across all age groups, while the importance of behavioral and socioeconomic factors varies with age. The research highlights that a substantial proportion of CVD cases and deaths could be attributed to these modifiable risk factors, suggesting that targeted preventive measures focusing on the most impactful factors for each age group could enhance strategies for CVD prevention and management. This approach marks a significant innovation in public health efforts, offering a more nuanced understanding of how age-specific risk factor modifications can inform more precise medical interventions.*

**基于思维链的单文档摘要生成示例2**:

**Source Document:**

*Title: Safety and efficacy of inhaled interferon-β 1a (SNG001) in adults with mild-to-moderate COVID-19: a randomized, controlled, phase II trial*

*With the emergence of SARS-CoV-2 variants resistant to monoclonal antibody therapies and limited global access to therapeutics, the evaluation of novel therapeutics to prevent progression to severe COVID-19 remains a critical need.....*

**GPT-3.5 Without COT:**

*In a randomized, controlled, phase II trial evaluating the safety and efficacy of inhaled interferon-β 1a (SNG001) in adults with mild-to-moderate COVID-19, it was found that SNG001 was safe with a non-statistically significant decrease in hospitalization for COVID-19 pneumonia compared to placebo. The trial, conducted between February 10 and August 18, 2021, showed no significant differences in adverse events, time to symptom improvement, or viral clearance between the SNG001 and placebo groups. However, there was a notable reduction in hospitalizations with SNG001 compared to placebo. The findings suggest that SNG001 may be a promising therapeutic option for preventing severe COVID-19 outcomes, although further research is needed to confirm its efficacy.*

**GPT-3.5 With COT:**

*The study evaluated the safety and efficacy of inhaled interferon-β 1a (SNG001) in adults with mild-to-moderate COVID-19 through a phase II randomized controlled trial. Adult outpatients with confirmed SARS-CoV-2 infection were randomized to receive either SNG001 or placebo for 14 days. The primary findings showed no significant differences in adverse events, time to symptom improvement, or viral RNA levels between the two groups. However, there was a trend towards fewer hospitalizations in the SNG001 group, with an 86% relative risk reduction compared to placebo. The study suggests that SNG001 was safe and may have potential benefits in reducing hospitalizations for COVID-19 pneumonia, although this reduction was not statistically significant in this trial.*

**Reference:**

*This study evaluates the safety and efficacy of inhaled interferon-β 1a (SNG001) for treating adults with mild-to-moderate COVID-19 through a phase II randomized controlled trial conducted between February 10 and August 18, 2021. Utilizing the ACTIV-2/A5401 platform, 220 adult*

CCL 2024

*outpatients with confirmed SARS-CoV-2 infection were randomized to receive either SNG001 or a placebo once daily for 14 days. The findings revealed no significant differences in adverse events, time to symptom improvement, or viral RNA quantification between the two groups. However, SNG001 was associated with a non-statistically significant reduction in hospitalizations, suggesting an 86% relative risk reduction compared to the placebo. This study introduces the potential of inhaled interferon-$\beta$ 1a as a novel therapeutic approach for COVID-19, highlighting its safety and a possible benefit in reducing hospitalizations, despite the need for further research to confirm its efficacy.*

**基于思维链的多文档摘要生成示例1：**

**Source Document:**

***Article1:*** *A Distillation Approach to Data Efficient Individual Treatment Effect Estimation*

.....

***Article2:*** *Estimation of Local Average Treatment Effect by Data Combination*

.....

***Article3:*** *Learning Instrumental Variable from Data Fusion for Treatment Effect Estimation*

.....

***Article4:*** *Treatment Effect Estimation with Data-Driven Variable Decomposition*

.....

**Qwen1.5 Direct:**

*This survey presents four innovative studies in the field of causal inference and treatment effect estimation, addressing various challenges in data efficiency and precision. "Data Efficient Individual Treatment Effect Estimator" (DEITEE) presents DEITEE, a method that combines a base model for confounding adjustment with a decision tree to select critical variables, reducing the number of required inputs while maintaining accuracy. It particularly shines in scenarios with limited data. The study on local average treatment effect (LATE) adaptation adapts a weighted least squares estimator to estimate LATE in merged datasets, relaxing positivity assumptions and outperforming IPW methods. The Group Instrumental Variable (GIV) approach, introduced in the Meta-EM algorithm, leverages representation learning to identify latent group labels without predefined IVs, overcoming limitations of traditional methods. Lastly, "Data-Driven Variable Decomposition" (D2VD) utilizes a regularized integrated regression model to separate confounders from adjustment variables, enhancing ATE estimation precision in high-dimensional data. These studies collectively demonstrate the advancement in causal inference techniques, addressing challenges in data handling and improving the accuracy of treatment effect estimation in various real-life settings.*

**Qwen1.5 Hierarchical:**

*Both sets of scientific articles delve into the realm of efficient and accurate treatment effect estimation in mixed observational datasets. The first set presents "Data Efficient Individual Treatment Effect Estimator" (DEITEE), a cutting-edge method that combines a base model for confounding adjustment with a decision tree to optimize predictions with limited data, particularly in scenarios with sparse individual information. This study complements the second one, which addresses non-compliance and missing treatment in a combined dataset by employing a weighted least squares estimator with estimated propensity-score-difference (PSD) to mitigate instability and relax positivity assumptions. Both works showcase their effectiveness through synthetic and real-world examples, highlighting their practicality in applications such as online-to-offline marketing and panel data analysis. The second set contributes to this area with its Group Instrumental Variable (GIV) method, a Representation Learning and Meta-EM algorithm that bypasses the need for predefined IVs by reconstructing latent source labels, relaxing linear*

CCL 2024

*regression assumptions and outperforming traditional summary IV methods. Meanwhile, Data-Driven Variable Decomposition (D2VD) proposes a regularization-based approach that separates confounders from adjustment variables in high-dimensional data, enhancing ATE estimation while incorporating orthogonality and sparsity constraints. These innovative methods demonstrate the potential of data-driven causal inference in various sectors, including social marketing, healthcare, and public policy, and highlight the importance of overcoming challenges in mixed observational datasets. Together, these studies emphasize the significance of developing efficient and robust techniques for estimating treatment effects in diverse settings.*

**Qwen1.5 Incremental:**

*This study introduces the Data-Driven Variable Decomposition (D2VD) algorithm, a novel approach to precise ATE estimation in observational studies by leveraging a regularized integrated regression model. D2VD effectively separates confounders from adjustment variables in high-dimensional data through orthogonality and sparsity constraints, outperforming traditional methods that treat all variables as confounders. It enhances accuracy and reduces variance in causal inference applications, particularly in areas like social marketing, healthcare, and public policy. In contrast, the research on Group Instrumental Variables (GIV) using representation learning and the Meta-EM algorithm in mixed observational datasets offers a breakthrough in causal effect estimation, eliminating the need for predefined IV candidates and learning latent source labels without linear regression assumptions. Although both studies demonstrate improved performance, GIV focuses on group differences and relaxes linear regression requirements, while DEITEE addresses limited data by combining a decision tree for confounding adjustment and individual treatment effect estimation. LATE estimation, another work, handles non-compliance and missing data with a weighted least squares estimator for Local Average Treatment Effect. These studies, along with their empirical validation, highlight their practical applications in various real-world scenarios and suggest potential for combined methods in complex causal inference tasks.*

**Reference:**

*In the realm of estimating treatment effects from observational data, recent advancements have introduced a variety of innovative methodologies to tackle the inherent challenges associated with data fusion, practical limitations of data acquisition, and the complexity of high-dimensional variables. Among these, the Data-Driven Variable Decomposition (D2VD) algorithm stands out by offering a solution to automatically separate confounders and adjustment variables in observational studies, thereby enabling more accurate treatment effect estimation with tighter confidence intervals, especially in scenarios characterized by high-dimensional data. This approach complements the Data Efficient Individual Treatment Effect Estimation (DEITEE) method, which reduces data requirements by identifying a minimal set of variables for accurate estimation, and a novel strategy that addresses the estimation of the local average treatment effect (LATE) in the presence of incomplete compliance and disjointed datasets through a weighted least squares estimator. Furthermore, the introduction of a method to overcome challenges posed by omitted source labels and unmeasured confounders in mixed datasets, through reconstructing the source label and modeling it as a Group Instrumental Variable (GIV) within a Meta-EM framework, signifies a significant leap forward. This framework estimates individual treatment assignment probability and infers treatment effect by mapping raw data into a representation space, estimating distribution differences, and modeling the GIV for different treatment assignment mechanisms through an alternating training strategy. Collectively, these studies not only expand the toolkit available for researchers and practitioners aiming to derive meaningful insights from observational data to inform treatment decisions and policy-making but also showcase the diversity of approaches to overcoming the practical challenges of data-related limitations and complex data structures in the estimation of treatment effects. The integration of the D2VD algorithm into this body of work highlights the importance of distinguishing between confounders and adjustment variables in enhancing the precision of treatment effect estimates,*


(c) 2024 中国中文信息学会计算语言学专业委员会

*thereby contributing to the broader field of causal inference with a nuanced understanding of variable impacts in observational studies.*

**基于思维链的多文档摘要生成示例2：**

**Source Document:**

***Article1:*** *Efficacy and safety of zapnometinib in hospitalised adult patients with COVID-19 (RESPIRE): a randomised, double-blind, placebo-controlled, multicentre, proof-of-concept, phase 2 trial*

.....

***Article2:*** *Safety and effectiveness of COVID-19 vaccines in patients with IgA nephropathy: a retrospective cohort study from the TriNetX global collaborative networks*

.....

***Article3:*** *Efficacy of a monovalent (D614) SARS-CoV-2 recombinant protein vaccine with AS03 adjuvant in adults: a phase 3, multi-country study*

.....

***Article4:*** *Safety and efficacy of inhaled interferon-β 1a (SNG001) in adults with mild-to-moderate COVID-19: a randomized, controlled, phase II trial*

.....

**GPT-3.5 Direct:**

*The studies examined various interventions for COVID-19 in different patient populations. One study investigated the safety and efficacy of zapnometinib in hospitalised COVID-19 patients, showing potential benefits in improving clinical severity scores, especially in non-Omicron variants. Another study focused on COVID-19 vaccination in IgA nephropathy patients, revealing reduced infection risks and normalized renal function post-vaccination. A third study assessed a SARS-CoV-2 recombinant protein vaccine, demonstrating efficacy in non-naïve individuals. Lastly, a study on inhaled interferon-β 1a in mild-to-moderate COVID-19 patients suggested a trend towards reduced hospitalizations. These findings collectively underscore the importance of diverse interventions in managing COVID-19, highlighting the need for further research to optimize treatment strategies and outcomes.*

**GPT-3.5 Hierarchical:**

*The collective findings from a diverse set of scientific articles shed light on various strategies in combating COVID-19. One study explored the safety and efficacy of zapnometinib in hospitalized COVID-19 patients, indicating potential benefits in improving clinical severity scores, especially in severe cases and non-Omicron variants, despite recruitment challenges. Concurrently, research on COVID-19 vaccination in IgAN patients revealed reduced infection risks post-vaccination, with a noted temporary decline in eGFR, emphasizing the importance of renal function monitoring. Another study delved into a monovalent SARS-CoV-2 recombinant protein vaccine with AS03 adjuvant, showcasing varied efficacy based on prior virus exposure, while a separate investigation on inhaled interferon-β 1a (SNG001) for mild-to-moderate COVID-19 suggested a potential decrease in hospitalizations. These studies collectively underscore the necessity for continued exploration of diverse approaches in combating COVID-19, with implications ranging from targeting specific pathways like Raf/MEK/ERK to monitoring renal health and addressing challenges in vaccine efficacy among naïve individuals, ultimately offering promising avenues for mitigating severe outcomes and advancing our understanding of effective interventions against the virus.*

**GPT-3.5 Incremental:**

*In a comprehensive exploration of COVID-19 interventions, a global Phase 3 trial assessed a monovalent SARS-CoV-2 recombinant protein vaccine with AS03 adjuvant, revealing varied*


(c) 2024 中国中文信息学会计算语言学专业委员会

*efficacy rates based on participants' serostatus. The vaccine demonstrated 15.3% efficacy in SARS-CoV-2 naïve individuals, 32.9% irrespective of serostatus, and 52.7% in non-naïve participants, emphasizing challenges in protecting naïve individuals against the D614 strain but showing promise for those previously exposed to the virus. Concurrently, research on zapnometinib in hospitalised COVID-19 patients highlighted potential benefits in improving clinical severity scores, particularly in severe cases and non-Omicron variants, with minimal adverse events. Additionally, a study focusing on COVID-19 vaccination in IgAN patients indicated reduced infection risks and complications post-vaccination, alongside a temporary eGFR decrease that normalized over time, stressing the importance of monitoring renal function in this population. The study evaluated the safety and efficacy of inhaled interferon-$\beta$ 1a (SNG001) in adults with mild-to-moderate COVID-19 through a phase II randomized controlled trial. Adult outpatients with confirmed SARS-CoV-2 infection were randomized to receive either SNG001 or placebo for 14 days. The primary findings showed no significant differences in adverse events, time to symptom improvement, or viral RNA levels between the two groups. However, there was a trend towards fewer hospitalizations in the SNG001 group, with an 86% relative risk reduction compared to placebo. The study suggests that SNG001 was safe and may have potential benefits in reducing hospitalizations for COVID-19 pneumonia, although this reduction was not statistically significant in this trial. These findings underscore the need for continued investigation into zapnometinib's therapeutic potential in COVID-19 treatment and the importance of optimizing outcomes through tailored interventions and vigilant monitoring.*

**Reference:**

*Recent research efforts have been directed towards identifying novel strategies to combat COVID-19 and mitigate its impacts, particularly in individuals with pre-existing health conditions. Among these, a pivotal study evaluated the safety, efficacy, and reactogenicity of a monovalent (D614) SARS-CoV-2 recombinant protein vaccine with AS03 adjuvant across 53 research centers in eight countries. This Phase 3 trial, randomized and placebo-controlled, aimed to assess the vaccine's effectiveness in preventing symptomatic COVID-19, revealing a nuanced efficacy profile: 15.3% in SARS-CoV-2 naïve participants, 32.9% regardless of initial serostatus, and 52.7% in previously exposed individuals, suggesting its potential utility as a booster despite challenges in protecting against new variants. The vaccine demonstrated an acceptable safety profile, aligning with the broader goal of identifying safe and effective COVID-19 interventions. Another innovative approach explored was the use of inhaled interferon-$\beta$ 1a (SNG001) in a phase II randomized controlled trial for treating adults with mild-to-moderate COVID-19. Although the study, conducted between February 10 and August 18, 2021, did not show significant differences in adverse events, time to symptom improvement, or viral RNA quantification between the treatment and placebo groups, SNG001 was associated with a non-statistically significant reduction in hospitalizations, suggesting an 86% relative risk reduction compared to placebo. This introduces the potential of inhaled interferon-$\beta$ 1a as a novel therapeutic approach. Additionally, zapnometinib, an oral MEK1/MEK2 inhibitor, was investigated for its potential immunomodulatory and antiviral effects in hospitalized adults with moderate to severe COVID-19. Despite early trial termination due to recruitment challenges, preliminary results indicated possible benefits in reducing clinical severity, especially in severe cases and those caused by non-Omicron variants, without significant adverse effects. Research utilizing the TriNetX collaborative networks also assessed the renal safety and effectiveness of COVID-19 vaccines in patients with Immunoglobulin A nephropathy (IgAN), finding a significant reduction in infection risk and complications, alongside a transient, reversible decline in renal function. Collectively, these studies underscore the critical importance of developing targeted therapeutic and preventive measures against COVID-19, emphasizing the need for ongoing research to refine treatment protocols and vaccination strategies in response to evolving viral variants and diverse patient health profiles.*

CCL 2024

**基于思维链的跨语言多文档摘要生成示例1：**

**Source Document:**

***Article1:*** *FiLM: Visual Reasoning with a General Conditioning Layer*

.....

***Article2:*** *Hierarchical ConViT with Attention-Based Relational Reasoner for Visual Analogical Reasoning*

.....

***Article3:*** *Learning the Dynamics of Visual Relational Reasoning via Reinforced Path Routing*

.....

***Article4:*** *Visual Concept Reasoning Networks*

.....

**ChatGLM3 Direct:**

本文综述了近年来在视觉推理领域的研究，这些研究主要集中在利用深度学习模型实现视觉推理。这些模型通常包含两个模块：感知模块和推理模块。感知模块通过卷积神经网络提取图像特征，而推理模块则利用这些特征进行逻辑推理。一些研究还探索了如何将条件化线性调制(FiLM)层引入视觉推理模型，以实现更复杂的推理任务。此外，一些研究还提出了基于注意力机制的模型，以更好地模拟人类视觉推理过程。这些模型通常包含多个模块，如卷积模块、转换器模块和注意力模块，以捕捉不同尺度的视觉信息。这些模型在解决视觉类比推理问题方面取得了良好的效果。此外，一些研究还探索了强化学习在视觉推理中的应用，以实现更准确和透明的视觉推理过程。总的来说，这些研究为视觉推理领域的发展提供了新的思路和方法，有望在未来的视觉理解和智能交互中发挥重要作用。

**ChatGLM3 chain1,3:**

本文综述了近年来在视觉推理领域的研究进展，包括使用特征线性调制层（FiLM）进行视觉推理，提出基于层次卷积视觉变换器（HCV-ARR）解决Raven's Progressive Matrix（RPM）问题，使用强化路径路由方法进行视觉关系推理，以及提出视觉概念推理网络（VCRNet）实现高级视觉概念之间的推理。这些研究都致力于开发更有效、更通用、更易于应用的视觉推理方法，以应对日益增长的数据和任务需求。这些方法在不同的视觉推理任务和数据集上表现出良好的性能，证明了这些方法的有效性和通用性。

**ChatGLM3 chain1,2,3:**

视觉概念推理网络（VCRNet）和两项最近的研究，其中一项引入了强化路径路由（RPR）和另一项引入了层次化卷积视觉变换器（HCV-ARR），这两项研究都为视觉推理和图像处理领域的发展做出了贡献。VCRNet提出了一种模块化的多分枝架构，通过推理高级视觉概念来捕获全局上下文，在多个识别任务上表现出优越性能，同时保持参数数量最小。另一方面，RPR利用强化学习在视觉图中学习多步路径，通过结合语言和历史注意力实现透明推理，在视觉关系推理任务上实现了最先进的准确度。HCV-ARR，一种结合卷积神经网络（CNNs）和视觉变换器的层次模型，强调了捕捉细粒度和全局上下文的重要性，通过有效地建模属性关系在RPM问题上超过了其他模型。这些研究展示了深度学习技术在处理复杂图像相关任务方面的灵活性和效率，每项研究都为视觉推理提供了独特的视角。尽管VCRNet专注于对高级概念进行推理，但RPR依赖于基于路径的推理和适应性，而HCV-ARR采用了一种结合卷积神经网络（CNNs）和变换器的层次方法。它们共同突显了视觉推理领域研究的不断发展和互补性。

**Reference:**

近期在神经网络调节和架构方面的进步显著提升了视觉推理能力，这是一个对深度学习模型传统上构成挑战的领域。特征线性调制（Feature-wise Linear Modulation, FiLM）层的引入标志着一个重大进步，提供了一种多功能的调节方法，使神经网络能够以提高的效率处理复杂的多步骤推理任务，如在CLEVR基准测试中所示。同时，分层ConViT与基于注意力的关系推理器（Hierarchical ConViT with Attention-based Relational Reasoner, HCV-ARR）的开发，解决了Raven's Progressive Matrices (RPMs)中视觉感知和逻辑推理的复杂性，将卷积块用于


(c) 2024 中国中文信息学会计算语言学专业委员会

低级视觉模式与变换器块用于高级图像语义相结合，并通过一个基于注意力的关系推理器增强推理能力，从而在RPM数据集上表现出色。此外，另一项研究提出通过将视觉关系推理概念化为路径路由任务来学习其动态性。这种方法采用加强的路径路由方法，从输入图像构建一个结构化的视觉图，并利用基于强化学习的模型来导航图中的路径（节点序列），由输入句子指导以推导出推理结果。在这些进步的基础上，视觉概念推理网络（Visual Concept Reasoning Networks, VCRNet）的引入采用了在卷积神经网络中用于视觉识别任务的分裂-转换-合并策略，以实现高级视觉概念之间的推理。通过将每个分支与一个视觉概念关联，并通过注意力模块派生紧凑的概念状态，VCRNet在通过基于图的交互更新概念状态后，适应性地调节局部描述符。这个模型通过分裂-转换-关注-交互-调节-合并阶段描述，展示了高度模块化架构在增强视觉识别任务方面的有效性，同时参数增加最小。总体而言，这些方法强调了整合多样化的神经网络机制——如特征调制、层次处理、基于注意力的推理和基于图的交互——以及像加强路径路由这样的创新方法，以应对视觉推理的多方面挑战的重要性。这一趋势突显了向更细腻和有效模型的转变，强调了推理的动态性和在模拟人类推理能力中建模认知过程的关键作用。

**基于思维链的跨语言多文档摘要生成示例2：**

**Source Document:**

***Article1:*** *Efficacy and safety of sintilimab plus albumin-bound-paclitaxel in recurrent or metastatic cervical cancer: a multicenter, open-label, single-arm, phase II trial*

.....

***Article2:*** *Comparative effectiveness of ustekinumab vs. vedolizumab for anti-TNF-naïve or anti-TNF-exposed Crohn's disease: a multicenter cohort study*

.....

***Article3:*** *Efficacy and safety of second-line therapy of docetaxel plus ramucirumab after first-line platinum-based chemotherapy plus immune checkpoint inhibitors in non-small cell lung cancer (SCORPION): a multicenter, open-label, single-arm, phase 2 trial*

.....

***Article4:*** *Fecal microbiota transplantation plus tislelizumab and fruquintinib in refractory microsatellite stable metastatic colorectal cancer: an open-label, single-arm, phase II trial (RENMIN-215)*

.....

**GPT-3.5 Direct:**

这些研究分别探讨了不同治疗方案在不同类型的癌症患者中的疗效和安全性。第一篇研究评估了Sintilimab联合白蛋白结合紫杉醇治疗宫颈癌的效果，显示出良好的抗肿瘤活性和可管理的毒性。第二篇研究比较了Ustekinumab和Vedolizumab在克罗恩病患者中的疗效，发现Ustekinumab在临床和客观结果上优于Vedolizumab。第三篇研究探讨了在非小细胞肺癌患者中，第一线免疫治疗后使用多西紫杉醇加拉姆西鲁单抗的二线治疗的疗效和安全性。最后一篇研究研究了在微卫星稳定性结直肠癌患者中，粪便菌群移植联合替雷利珠单抗和富奎替尼的疗效，显示出改善生存和可管理的安全性。这些研究为不同癌症类型的治疗提供了新的治疗选择，并强调了进一步进行大规模随机对照试验的重要性。

**GPT-3.5 chain1,3:**

这些研究均探讨了不同治疗方案在晚期癌症患者中的疗效和安全性。一项研究评估了在复发或转移性宫颈癌患者中将sintilimab与白蛋白紧密结合紫杉醇联合应用的效果，显示出潜在的治疗前景。另一项研究比较了ustekinumab和vedolizumab在克罗恩病患者中的效果，结果表明ustekinumab可能是更有效的选择。还有一项研究探讨了docetaxel加ramucirumab作为非小细胞肺癌患者的二线治疗选择，显示出良好的抗肿瘤活性和可管理的安全性。最后一项研究则研究了将粪便菌群移植与tislelizumab和fruquintinib结合应用于难治性微卫星稳定性结直肠癌患者的疗效，强调了肠道菌群在免疫治疗疗效中的重要作用。这些研究为癌症治疗领域提供了新的治疗选择和见解，需要进一步的临床验证和研究。

CCL 2024

**GPT-3.5 chain1,2,3:**

在一系列探索不同癌症治疗方案的疗效和安全性的多样化研究中，涌现出了创新的方法。一项研究调查了将粪便微生物移植（FMT）与替雷利珠单抗和富奎替尼结合在难治性微卫星稳定（MSS）转移性结直肠癌（mCRC）患者中的应用，显示出有希望的结果，中位无进展生存期（PFS）为9.6个月，并且在反应者中发现了肠道微生物组成的改变，突显了肠道菌群在免疫治疗疗效中的作用。另一项研究关注了多西紫杉醇加拉莫西单抗作为非小细胞肺癌（NSCLC）患者的二线治疗，显示出34.4%的客观缓解率和可控的不良事件，而另一项研究探讨了辛地利单抗和白蛋白结合紫杉醇用于复发性或转移性宫颈癌，显示出有希望的缓解率和可耐受的副作用。此外，一项比较克唑替尼单抗和维多利珠单抗在克罗恩病患者中的研究发现，对于TNFi未曾接受和TNFi曾接受的个体，克唑替尼单抗在第26周和第52周具有更优越的临床结果。这些研究共同呈现了各种癌症的宝贵新治疗选择，强调了个性化治疗策略的重要性，并为在临床实践中进一步探索提供了有希望的途径。

**Reference:**

近期的临床研究为慢性疾病的创新疗法开辟了道路，展示了针对性治疗在各种癌症疾病中的潜力。其中，日本进行的第二阶段SCORPION试验在非小细胞肺癌（NSCLC）的治疗中显示出了希望，该试验通过将多西他赛与雷莫芦单抗结合作为二线疗法，达到了34.4%的客观响应率（ORR），且安全性可控。同样，将针对PD-1的抗体西妥昔单抗与白蛋白结合型紫杉醇用于复发或转移性宫颈癌的组合，显示出了显著的抗肿瘤活性，ORR为44.4%，疾病控制率为88.9%，尽管近半数参与者出现了可管理的高级别不良事件。在胃肠疾病领域，一项多中心队列研究强调了在治疗克罗恩病方面，乌司他单抗相比于维多利单抗的优越效果，分别在26周和52周时达到了临床和无需激素的缓解，且安全性相当。在这一系列针对性疗法中，研究“粪便微生物群移植加上替雷利珠单抗和复仑替尼治疗难治性微卫星稳定型转移性结直肠癌”引入了一种针对难治性微卫星稳定型转移性结直肠癌（mCRC）的创新治疗方案，结合了粪便微生物群移植（FMT）与抗PD-1和抗血管生成疗法。该试验报告了中位无进展生存期（PFS）为9.6个月，总生存期（OS）为13.7个月，且安全性可控，凸显了肠道微生物组在提高免疫疗效中的重要作用。这些研究共同强调了结合针对性疗法以改善晚期癌症和慢性疾病患者结果的潜力，暗示了向更个性化和有效的治疗策略转变。

## C 数据集示例

### C.1 人工智能领域

下面簇中包含4篇文档:

**Article1:**

Title: FiLM: Visual Reasoning with a General Conditioning Layer

（该文档共904个单词）

**Article2:**

Title: Hierarchical ConViT with Attention-Based Relational Reasoner for Visual Analogical Reasoning

（该文档共1192个单词）

**Article3:**

Title: Learning the Dynamics of Visual Relational Reasoning via Reinforced Path Routing

（该文档共918个单词）

**Article4:**

Title: Visual Concept Reasoning Networks

（该文档共693个单词）

**跨语言多文档摘要:**

近期在神经网络调节和架构方面的进步显著提升了视觉推理能力，这是一个对深度学习模型传统上构成挑战的领域。特征线性调制（Feature-wise Linear Modulation, FiLM）层的引入


(c) 2024 中国中文信息学会计算语言学专业委员会

标志着一个重大进步，提供了一种多功能的调节方法，使神经网络能够以提高的效率处理复杂的多步骤推理任务，如在CLEVR基准测试中所示。同时，分层ConViT与基于注意力的关系推理器（Hierarchical ConViT with Attention-based Relational Reasoner, HCV-ARR）的开发，解决了Raven's Progressive Matrices (RPMs)中视觉感知和逻辑推理的复杂性，将卷积块用于低级视觉模式与变换器块用于高级图像语义相结合，并通过一个基于注意力的关系推理器增强推理能力，从而在RPM数据集上表现出色。此外，另一项研究提出通过将视觉关系推理概念化为路径路由任务来学习其动态性。这种方法采用加强的路径路由方法，从输入图像构建一个结构化的视觉图，并利用基于强化学习的模型来导航图中的路径（节点序列），由输入句子指导以推导出推理结果。在这些进步的基础上，视觉概念推理网络（Visual Concept Reasoning Networks, VCRNet）的引入采用了在卷积神经网络中用于视觉识别任务的分裂-转换-合并策略，以实现高级视觉概念之间的推理。通过将每个分支与一个视觉概念关联，并通过注意力模块派生紧凑的概念状态，VCRNet在通过基于图的交互更新概念状态后，适应性地调节局部描述符。这个模型通过分裂-转换-关注-交互-调节-合并阶段描述，展示了高度模块化架构在增强视觉识别任务方面的有效性，同时参数增加最小。总体而言，这些方法强调了整合多样化的神经网络机制——如特征调制、层次处理、基于注意力的推理和基于图的交互——以及像加强路径路由这样的创新方法，以应对视觉推理的多方面挑战的重要性。这一趋势突显了向更细腻和有效模型的转变，强调了推理的动态性和在模拟人类推理能力中建模认知过程的关键作用。

### C.2 医学领域

下面簇中包含3篇文档:

**Article1:**

Title: Estimates of the global burden of non-Hodgkin lymphoma attributable to HIV: a population attributable modeling study

（该文档共383个单词）

**Article2:**

Title: The burden of non-communicable diseases among people living with HIV in Sub-Saharan Africa: a systematic review and meta-analysis

（该文档共267个单词）

**Article3:**

Title: Burden and risk factors of chronic obstructive pulmonary disease in Sub-Saharan African countries, 1990-2019: a systematic analysis for the Global Burden of disease study 2019

（该文档共355个单词）

**跨语言多文档摘要:**

近期研究揭示了撒哈拉以南非洲（SSA）面临的紧迫健康挑战，重点关注了艾滋病毒感染者（PLHIV）中非传染性疾病（NCDs）的负担以及慢性阻塞性肺疾病（COPD）构成的日益严峻的公共卫生挑战。一项综合性的元分析揭示，2019年全球新增非霍奇金淋巴瘤（NHL）病例中有6.92%可归因于HIV，东部和南部非洲以及东欧和中亚的流行率显著，凸显了PLHIV中NHL风险增加。此外，一项系统评价和元分析发现，撒哈拉以南非洲的PLHIV中高血压、抑郁症、糖尿病和其他NCDs的比率值得注意，强调了加强卫生系统和综合医疗服务的必要性。与此同时，一项分析1990年至2019年全球疾病负担研究数据的研究发现，SSA的COPD流行率急剧增加，2019年的流行病例达到1030万，自1990年以来增加了117%。该研究指出，固体燃料的家庭空气污染是COPD的主要风险因素，尤其是在中部撒哈拉以南非洲，通过残疾和过早死亡显著地加剧了疾病负担。这两组发现强调了制定全面健康策略的迫切需要，这些策略旨在应对HIV等传染病和NCDs（包括COPD）日益增长的影响的双重挑战，以改善SSA的整体健康结果。这种综合方法与联合国艾滋病规划署的"90-90-90"快速行动目标相一致，并为该地区针对性的公共卫生干预和政策制定提供了关键见解，凸显了传染病和NCDs在塑造公共卫生优先事项中的相互关联性。


(c) 2024 中国中文信息学会计算语言学专业委员会

# 面向语言学习者的跨语言反馈评语生成方法

**安纪元[1,2]，朱琳[1,2]，杨尔弘[1,2*]**
[1]北京语言大学国家语言资源监测与研究平面媒体中心，北京　100083
[2]北京语言大学信息科学学院，北京　100083
jiyuanan.blcu@gmail.com

## 摘要

反馈评语生成任务旨在为语言学习者的产出提供纠偏及解释性的评价，促进学习者写作能力的发展。现有研究主要聚焦于单语的反馈评语生成，如为英语学习者提供英文反馈评语，但这忽略了非母语学习者可能面临的理解障碍问题，尤其当评语中存在陌生的语言知识时。因此，本文提出跨语言反馈评语生成任务（CLFCG），目的是为语言学习者生成母语的反馈评语。本研究构建了首个英-中跨语言反馈评语生成数据集，该数据集包含英语学习者产出的语句与相应的中文反馈评语，并探索了基于流水线的预训练语言模型引导增强生成方法，将修正编辑、线索词语和语法术语等作为输入的附加信息，引导和提示生成模型。实验结果表明，附加引导信息的预训练语言模型流水线方法在自动评估（BLEU: 50.32）与人工评估（Precision: 62.84）上表现良好。本文对实验结果进行了深入分析，以期为跨语言反馈评语生成任务提供更多见解。

**关键词：** 智能辅助语言学习；反馈评语生成；跨语言文本生成；预训练语言模型

# Cross-Lingual Feedback Comment Generation for Language Learners

**Jiyuan An[1,2], Lin Zhu[1,2], Erhong Yang[1,2*]**
[1]National Language Resources Monitoring and Research Center for Print Media, Beijing Language and Culture University, Beijing
[2]School of Information Science, Beijing Language and Culture University, Beijing
jiyuanan.blcu@gmail.com

## Abstract

The task of generating feedback comments is designed to furnish corrective and interpretative evaluations for the outputs of language learners, thereby facilitating the enhancement of their writing competencies. Predominantly, existing studies have concentrated on monolingual feedback generation, such as providing feedback in English for learners of English. However, this approach neglects the comprehension barriers that non-native speakers may encounter, particularly when the feedback incorporates unfamiliar linguistic elements. To address this issue, this paper introduces the task of Cross-Linguistic Feedback Comment Generation (CLFCG), which aims to produce feedback in the learner's native language. This research develops the inaugural English-Chinese cross-linguistic feedback comment generation dataset, comprising sentences

* 通讯作者
**基金项目**：国家语委科研项目（ZDA145-17）；教育部人文社会科学研究一般项目（23YJCZH264）；中央高校基本科研业务费（北京语言大学梧桐创新平台，21PT04）
©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

crafted by English learners paired with corresponding feedback in Chinese. The study explores a methodology based on a pipeline of pre-trained language models, enhanced by guided generation techniques that incorporate modification edits, cue words, and grammatical terminology as supplementary input to steer and prompt the generative model. The experimental outcomes demonstrate that the enhanced pre-trained language model pipeline with additional guiding information yields robust performance, achieving a BLEU score of 50.32 and a Precision of 62.84 in automatic and human evaluations respectively. An in-depth analysis of these results is presented, offering substantive insights for the advancement of cross-linguistic feedback comment generation tasks.

**Keywords:** Intelligent Computer-Assisted Language Learning , Feedback Comment Generation , Cross-Lingual Text Generation , Pretrained Language Models

## 1 引言

智能辅助写作是计算机智能辅助语言学习（ICALL）中的一个重要研究领域，其主要目的在于帮助语言学习者纠正写作中的错误，以提高他们的语言运用能力。近年来，自动语法错误纠正任务（GEC）受到了广泛的研究关注。该任务通过识别学习者文本中存在的错误，生成修正后的正确语句，旨在减轻语言教师的负担，使语言学习过程更加即时和便捷。然而，自动语法错误纠正任务通常只能告诉学习者“改哪里”和“如何改”，而忽略了最为本质的问题——“为什么要修改”。这种方法可能导致语言学习者在一知半解中修改了本次的错误，但是在未来的表述中出现同样的错误。反馈评语生成任务正是为了解决这一问题而设计的。反馈评语是帮助作者（学习者）提高写作技巧的字符串，它不仅提供关于语法错误的具体反馈，还可以包括关于话语、结构和内容等其他方面的评价。这种反馈既可能是对当前写作的具体指导，也可能是对作者的鼓励。总之，为语言学习者的写作提供针对性的建议或解释性评语，对于提高他们的语言运用水平是极其有益的。

在本研究之前，由于平行训练和评估数据的可获得性，有关反馈评语生成任务的研究主要集中于单一语言环境。然而，这些研究往往忽略了一个重要的实际问题，即生成的评语可能包含对语言学习者来说陌生的词汇。换言之，这更适合为希望提高写作水平的英文母语者生成反馈评语。如图1 所示，如果以汉语为母语的英语学习者想要知道错误的原因，相比仅提供结果的英文语法错误纠正和英文反馈评语，中文反馈评语更易于被理解和接受。

因此，本文提出了一项新颖的跨语言反馈评语生成（TLFCG）的新任务，并构建了首个针对英文学习者语句采用中文反馈评语的跨语言反馈评语数据集。在数据集的构建过程中，本研究利用现有的反馈评语生成数据资源，在构建语法术语翻译映射表的基础上，利用大语言模型以无监督的方式进行翻译，并通过人工标注对结果进行修正。为了实现跨语言反馈评语生成，本研究尝试将修正编辑、线索词语及语法术语等反馈评语中的关键信息整合入流水线，作为生成模型输入的附加信息，从而起到对输出的引导和提示作用，提高模型生成效果。研究结果表明，引入不同的附加信息会对生成效果产生显著影响，与直接微调的基线模型相比，本文所提方法能够生成更高质量的反馈评语。此外，我们对实验结果进行了深入分析，以期为今后的相关研究提供参考。本文的主要贡献有：构建首个针对中文母语者学习英语外语的反馈评语语料库；探索预训练语言模型和大语言模型跨语言反馈评语生成的能力；分析添加不同附加信息对预训练语言模型生成反馈评语效果的影响。

## 2 相关工作

智能辅助写作是计算机智能辅助语言学习领域的一个重要研究方向。由于传统的语法错误纠正任务缺乏可解释性，无法向语言学习者提供修改的原因，因此Nagata(2019)提出反馈评语生成任务，旨在为语言学习者提供关于错误原因的解释性说明，从而帮助语言学习者提升写作水平。Nagata(2020)发布了基于亚洲英语学习者国际语料库网络（ICNALE）(Ishikawa, 2013)标注反馈评语的细节，并公开了部分标注结果。该数据集采用英语（标注结果被弃用）和日语标注反馈评语，涵盖了多种错误类型。其中，反馈评语中的语法术语和引文采用两个


(c) 2024 中国中文信息学会计算语言学专业委员会




图 1: 跨语言反馈评语生成任务

特殊符号进行标记。语法术语使用“<”和“>”标记（如<不及物动词>）以便于学习者对照语法书中相应的语法项目。而引用符号“<<”和“>>”用于表明其中的单词是从原始语句中引用的（如<<agree>>），这使得反馈意见更加灵活和具体。在INLG 2022针对语言学习者反馈评语生成的共享任务（GenChal）中，Nagata(Nagata et al., 2021)进一步定义了反馈评语生成任务，即输入是学习者语句和需要进行评语的位置索引，输出是针对该位置的具体反馈评语。同时，Nagata公开了一个基于上述数据集翻译而来的、专注于介词使用错误类型的英文单语言数据集(Nagata et al., 2023)。

在反馈评语生成方法的研究中，研究者尝试了不同的模型架构和训练策略。Hanawa(2021)对基于检索、生成以及检索与编辑结合的三种方法进行了初步探索。研究发现，检索与编辑结合的方法虽然能够对检索结果进行修改，但其过度编辑的问题导致生成效果不佳；基于检索的方法虽稳定，却因仅限于检索已有评语库中的匹配项而缺乏灵活性；相较之下，简单生成的方法取得了最佳的性能，其使用序列到序列模型直接生成评语，能够提供更多样化的反馈，但其生成评语准确性和可用性仍是一个挑战。

为了提高模型的泛化能力和性能，研究者还探索了多种数据增强技术。Babakov(2022) 通过使用依存句法分析对原始学习者语句进行了裁剪，并采用语言模型GPT-Neo对剪裁后的部分进程扩展以生成伪数据。而Behzad(2022) 则采取了一种不同的伪数据生成方式，该方法通过在ICNALE语料库中标注未用于训练和验证的其他文章中的介词使用类型错误，从而扩大了数据规模。

为了在生成过程中充分利用和参考已有的数据，Ihori(2023) 提出了基于检索的生成方法，该方法包括三个主要模块：检索模块、屏蔽模块和生成模块。首先，检索模块从训练数据中检索与输入学习者句子最相似的实例。然后，屏蔽模块将屏蔽检索到的反馈评语中与输入句子不太相符的词汇。最后，生成模块依据输入句子及已屏蔽的反馈评语来生成最终的反馈评语。检索和掩蔽模块基于BERT模型，而生成模块使用预先微调的T5模型。另一方面，Jimichi(2023)采用了预训练的T5模型作为生成器，并使用RoBERTa作为分类器来获取名词、介词等语法术语标签。这些预测出的语法术语标签被用作生成模型中的一个额外信息源。

此外，在语法纠错任务中，关于可解释性的研究也并未止步。Fei(2023) 认为对于语法错误纠正任务的可解释性而言，错误的原因（线索词语）及其对应的错误类型是解释错误的两个关键因素。为了通过解释来增强GEC模型，该研究引入了一个配有线索词语和语法错误类型标注的大型数据集——EXPECT，并基于此数据集提出了结合语法分析和错误修正机制的两个基线模型。


(c) 2024 中国中文信息学会计算语言学专业委员会

## 3 数据

本研究利用现有反馈评语生成数据资源，构建了首个英-中跨语言反馈评语生成数据集。本章节将介绍数据来源和数据集构建的详细过程。

### 3.1 数据来源

现有反馈评语生成数据集，主要有基于亚洲英语学习者国际语料库网络（ICNALE）标注的数据集和INLG 2022 针对语言学习者反馈评语生成共享任务（FCG GenChal）的数据集（Nagata，2020）。前者是从ICNALE语料库中抽取了中国大陆、中国台湾、日本、韩国、泰国、印度尼西亚等国家和地区的1,194篇文章，使用日语对其标注了反馈评语，其中400篇来自中国的作者，约占33.5%。虽然这也是一个跨语言的数据集，但该数据集并未得到实际的使用。而后者是对前者中关于介词使用错误类型进行筛选并将评语语言翻译为英文后的结果，这也是目前唯一被广泛用于反馈评语生成任务的数据集。语料库规模如表1所示。

| 数据集名称 | | 学习者语句数量 | 总词语数量 | 反馈评语数量 |
|---|---|---|---|---|
| ICNALE 数据集 | | 17,938 | 300,289 | 10,463 |
| 共享任务数据集 | 训练集 | 4,868 | 110,906 | 4,868 |
| | 开发集 | 170 | 3,142 | 170 |
| | 测试集 | 215 | 4,446 | 215 |

表 1: 反馈评语数据集规模

### 3.2 数据标注

#### 3.2.1 数据清洗

本研究首先对原始数据中存在错误进行了人工修改，包括以下几个方面:

**错误符号** 语法术语符号、原句引用符号和引号的使用需要满足匹配关系。在这一步骤中，包括非法语法术语（例如<verb >>→ <verb>）、非法原句引用（例如<<couple>→ <<couple>>）和非法引号（例如，'of" → "of"）等错误已得到纠正。

**符号混淆** 语法术语符号和原句引用符号分别为“<...>”和“<<...>>”，但是在原始评语数据集中，二者存在一部分混合使用的情况，例如语法数据“<verb>”被标记为“<<verb>>”。本文通过人工逐条筛查的方式修正了这类错误。

**非法索引** 目标索引的开始和结束必须分别对应于错误单词的开始和结束，而不能包含单词之间的空格。例如，对于输入“It is fun to me .”，10:12 的索引是正确的，而9:12 的索引是非法的。本文通过将字符级别的索引自动转化为单词级别的索引解决了这一问题。

**字符编码** 其他一些更为少见的错误，如语法错误和非ASCII 字符的使用（例如全角符号'Ａ'和半角符号'A'）也在数据清洗中被修改。

#### 3.2.2 评语翻译

对于生成学习者母语的跨语言反馈评语，一个最自然的想法是在现有单语言反馈评语生成结果的基础上挂载一个翻译模块。因此，本研究首先尝试使用商业机器翻译模型进行英译中的评语翻译。然而，根据本文的测试结果，由于反馈评语中包含表示语法术语和学习者语句引用的特殊符号（即“<”、“>”、“<<”和“>>”），传统机器翻译模型通常无法对其进行处理，传统机器翻译模型通常无法处理含有表示语法术语和学习者语句引用的特殊符号（如“<”、“>”、“<<”和“>>”）的反馈评语，因而导致反馈评语中包含的语法术语无法被正确地翻译的，从而进一步造成学习者的困惑。外挂机器翻译模型的方法存在对语法术语翻译不准确（例如将<intransitive verb>译为<in及物动词>）、无法理解评语中特殊符号（部分翻译或过度翻译）等问题，这也间接证明了对跨语言反馈评语生成进行研究的必要性。基于此，本文放弃直接使用翻译模型的方法构建跨语言的反馈评语生成任务数据集，转而采用人工校对基于语法术语对照的大语言模型辅助翻译方法。

**语法术语翻译** 在原始数据中，相同的语法点存在多种表述方式（例如“determiner”和“qualifiers”均表示限定词），这是与反馈评语生成任务的定义相悖的，既不利于语言学习


(c) 2024 中国中文信息学会计算语言学专业委员会

者的学习和记忆，也不利于构建语法知识点对应的详细解释。因此，本文参考一线大学英语授课教师的建议，通过人工翻译和整理，将原有的2,202个日文语法术语和348个英文语法术语全部对应至1,745个中文语法术语，减少其中相同描述，并使其更加符合中国的英语教学实际需要。

**机器翻译辅助** 本文基于上述构建的语法术语翻译对照，使用GPT-4-Turbo模型对英文反馈评语进行翻译，以解决使用传统机器翻译模型时对于特殊符号之间内容处理有误的问题。为了使其生成的翻译结果符合预期，本文在Prompt中加入了对翻译要求的描述和一些输入输出示例作为提示。除此之外，本研究基于上述构建语法术语对应集，将语法术语翻译和原句引用作为替换操作，以Prompt的形式输入GPT模型，从而在翻译过程中保留对于语法术语和原句引用的特殊符号，所用Prompt如表2所示。同时，为了使模型拥有更好的效果，本文在调用API时使用了System Prompt: "You are ChatGPT, a large language model trained by OpenAI, based on the GPT-4 architecture.\nKnowledge cutoff: 2023-12\nCurrent date: 2024-03-10"。

| 日-中评语翻译Prompt | 英-中评语翻译Prompt |
|---|---|
| 请将下面的日语句子翻译成中文，不要保留任何日语。在翻译中，请使用以下给定的词语替换：<br>{term_str}；{reference_str}。直接输出翻译后的中文句子，不要包含任何其他内容。<br>日语句子：{feedback}<br>中文句子： | 将下面的英文句子翻译成中文。在翻译中，请使用以下给定的词语替换：<br>{term_str}；{reference_str}。直接输出翻译后的中文句子，不要包含任何其他内容。<br>英文句子：{feedback}<br>中文句子： |
| term str = {非母语语法术语} ->{中文语法术语}<br>reference str = {原句引用} ->{原句引用} | |

表 2: 使用GPT-4模型翻译反馈评语所用Prompt

**人工校对** 为进一步提升中文反馈评语的质量，本研究招募了10名以中文为母语所学专业为日语或专业水平N2以上的本科生和研究生，在GPT-4模型翻译结果的基础上，对中文反馈评语进行校对。其主要任务是：(1)检查翻译结果是否符合中文的表达习惯；(2)检查翻译结果与原始反馈评语是否表达相似的含义；(3)检查翻译后的反馈评语是否包含与原始反馈评语相同的语法术语、原句引用等；(4)检查翻译后的反馈评语中包含的特殊符号是否满足两两匹配关系以及其它格式是否正确。我们组织标注者参加反馈评语生成任务翻译的培训会议，并进行了对20条翻译结果的校对试标注，标注者完成试标注后可以进入正式的标注任务，我们对标注结果进行最终审查。为了评估模型在实际使用场景下的效果，我们邀请了两位英语教师来校对测试集。

## 4 实验

基于前一章节构建的跨语言反馈评语数据集，本研究首先对任务定义进行了明确，然后分别采用附加引导的流水线预训练语言模型方法和少量样本提示的大语言模型方法进行了实验，并对生成结果的评估指标进行了详细介绍。

### 4.1 任务定义

跨语言反馈评语生成通常指在对于给定的学习者语句自动生成学习者母语中的反馈评语。在本研究中，输入数据是一个由英文语句及其错误位置索引组成的集合，输出则是针对指定错误位置的反馈评语。其中错误位置索引是一个整数区间，用于指出学习者语句中需要进行解释性说明的具体位置。所生成的反馈评语旨在帮助作者（语言学习者）提升写作技能，通常包含对语法错误的点评，同时也可能涉及话语、结构及内容方面的建议。这些建议不仅能改善当前的写作，也可能包含对作者的鼓励或赞美。

由于介词使用对于非母语学习者存在一定难度，其中含有介词区分、固定搭配、特殊用法和习惯用语等难点。并且介词使用错误类型在现有数据集中拥有最大的数据量，在所有错误类型中占有最高的比例。因此，本研究以介词使用错误类型的跨语言反馈评语生成作为研究起


(c) 2024 中国中文信息学会计算语言学专业委员会

点。而Babakov（2022）和Behzad（2022）等已经证明，对于低资源的反馈评语类型，可以通过数据增强的方法显著提升模型的性能。因此，对于数据量较少的错误类型，可以通过数据增强的方式扩展数据规模，而后采用与介词错误相同的方法生成反馈评语。

## 4.2 方法

### 4.2.1 预训练语言模型方法

反馈评语生成任务是将语言学习者撰写的输入文本转换为解释语法规则的另一文本的一种序列到序列生成任务，这意味着在其他生成式任务中被证明有效的方法（如复制机制）可能同样有益于这一任务（Nagata，2023）。

通过以下示例，我们发现一条反馈评语通常包含四个特殊的部分，即错误引用、修正编辑（Correction Edit，CE）、线索词语（Evidence Word，EV）和语法术语（Grammatical Terms，GT）。

***Input***: *The small steps will lead to a complete ban of smoking not only at restaurants , but also at any other indoor places . 90:92*

***Output***: <介词> <<*at*>>可以与<<*place*>>一起使用来指示某事发生的地点，但更常见的是使用*'in'*代替。

反馈评语通常引用输入文本中出现的单词和短语，错误引用“at”已经在输入中被标记，而未被标记线索词语“places”也会出现在原始语句中。同时，在反馈评语中使用符号“<”和“>”标记的语法术语，以及引号“”之间的修正编辑，对提升反馈评语生成的效果也可能是有益的。如果这些信息被附加在生成模型输入中，即可便于该模型引用源文本中的这些片段，从而降低生成任务的难度。

对于生成模型而言，同时识别原句中的错误引用，并预测修正编辑、线索词语和语法术语，然后将它们组织成一个完整的句子，是一项极具挑战性的任务。因此，我们考虑将任务细分为多个步骤，采用流水线的方式处理。首先预测修正编辑、线索词语和语法术语等信息，然后基于这些附加信息生成最终的反馈评语。这样的步骤化分解有助于模型更好地处理复杂信息，从而提高反馈评语的准确性。对于跨语言反馈评语生成任务，我们基于前文构建的数据集，选择采用多语言预训练语言模型mBART(Tang et al., 2020)和mT5(Xue et al., 2021)作为反馈评语生成模型进行微调。我们的任务流程如图2所示。

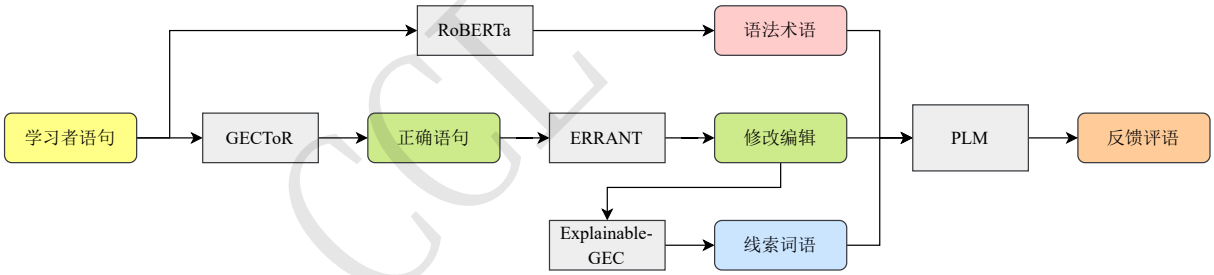



图 2: 跨语言反馈评语生成任务流程

**直接微调** 基于前文构建的数据集，我们直接微调了mBART和mT5预训练语言模型并作为跨语言反馈评语生成任务的基线。由于预训练语言模型无法有效处理输入中的索引位置，因此我们将原始输入中基于字符的错误位置索引修改为基于单词的位置索引，并在需要评语的位置起始和结束分别添加符号“[”和“]”作为标记。

**修正编辑** 由于约34%的反馈评语中包含了对于错误位置的修改建议，因此我们使用GECToR(Omelianchuk et al., 2020)模型对语料中的所有语句进行了自动修正，以得到语法正确的语句。然后，为了进一步得到错误位置的修改编辑，我们使用了ERRANT(Bryant et al., 2017)工具，提取原始语句和修正语句之间的编辑操作，对于编辑操作在待评语索引范围内的编辑进行保存。针对上例中的修正语句为“The small steps will lead to a complete ban of smoking not only at restaurants , but also in any other indoor places .”，保留的ERRANT结果为：“Orig: [11, 12, 'at'], Cor: [11, 12, 'in'], Type: 'R:PREP'”，即将介词‘at’替换为‘in’。


(c) 2024 中国中文信息学会计算语言学专业委员会

**线索词语** 线索词语通常可以提供有关语法错误发生原因的线索或指示。例如，在处理介词错误时，介词本身及其周围的名词短语常常是关键线索。通过在生成模型训练中加入这些线索词语的标注，可以显著提高模型的错误识别和修正准确率。我们基于已经获得的修正编辑，直接使用了Fei(2023)公开的线索词语生成模型，得到线索词语。

**语法术语** 我们注意到几乎所有反馈评语中均包含有语法术语（GT），如果在生成反馈评语之前，将预测出可能使用的语法术语附加到输入中，可能会显著提高模型的性能。在上一章节中，我们已经构建出了完整的语法术语表。由于每条反馈评语中可能使用了多个语法术语，因此可将这视为一个多标签分类任务。因此，我们借鉴了Jimichi(2023) 提出的基于RoBERTa(Liu et al., 2019)附加一个线性层的语法术语预测模型，并在中文数据集上进行训练，同样仅保留出现频率前十的语法术语作为附加信息。

我们将上述三种附加提示分别追加在以制表符"|"作为分割的原始学习者语句之后作为输入，并尝试将上述三者分别两两组合或全部组合作为输入，分别微调了mBART和mT5预训练语言模型。

#### 4.2.2 大语言模型方法

鉴于大语言模型在语言理解和指令遵循方面的强大能力，以及其在通用领域和众多自然语言处理任务中的优异表现，本文评估了大语言模型（使用GPT-3.5和GPT-4，采用few-shot方法）在跨语言反馈评语生成任务上的效果，并初步探索了提升其在此任务上表现的方法。

本研究重点关注于构建有效的Prompt，以充分利用大语言模型的指令遵循能力。为了使大语言模型更好地理解和适应跨语言反馈评语生成任务，我们在详细描述任务的基础上，尝试了zero-shot、one-shot和few-shot等不同的Prompt构造策略。通过不断优化Prompt内容，我们从数十个版本中筛选出了效果最佳的Prompt。我们要求大语言模型扮演中国英语教师的角色，使其与对英语外语教师的行为对齐。此外，我们分别添加了对于错误位置进行添加、修改、删除等三种类型的示例，以使大语言模型更好的理解跨语言反馈评语任务描述、需求和输出模式。详细的实验设置请见附录A。

### 4.3 评价

#### 4.3.1 自动评估指标

**表层相似度** 与前人关于反馈评语生成的研究相同，本文使用BLEU (Bilingual Evaluation Understudy) (Papineni et al., 2002)作为评估结果表层相似度的指标。BLEU 最初被用于评估机器翻译的质量，其通过统计模型输出与一组参考之间的N-gram 匹配的次数来可以反映二者之间的接近程度。本研究直接使用Ryo Nagata（2021）在共享任务种提供评估代码工具来计算BLEU分数。

**语义相似度** 对于语义相似度的评估，本研究采用了BERTScore (Zhang et al., 2020)。BERTScore 利用预训练的语言模型BERT 来计算生成反馈与参考反馈之间的语义相关性。具体而言，BERTScore 通过计算反馈文本中每个词语的BERT 嵌入向量与参考文本中最相似词语的余弦相似度，来评估整体的语义匹配度。这种方法不仅关注词语的表层匹配，而且更加重视语义的深层次相似性，能够更全面地评估生成文本的质量。

#### 4.3.2 人工评估指标

在人工评估方法中，我们采用了Nagata(Nagata et al., 2021)提出的评估框架。具体来说，三位独立的评价者在盲审的情况下，使用0至2的评分系统基于特定评估标准对每条生成的反馈评语进行评估：完全正确（2）、部分正确（1）或不正确（0）。其中，**完全正确（2）**指反馈评语不仅包含了与参考内容相似的信息，而且未包括任何与错误无关的内容。即便反馈中包含了参考内容未提及的相关信息，只要这些信息直接关联到错误，也被认定为完全正确。**部分正确（1）**的评分意味着反馈评语基本准确，但需要简单编辑以提高准确度；例如，如果反馈正确指出了句子的错误，且只需修改几个词便能让内容更准确，则归于此类。而对于那些与参考内容完全无关，未能指出错误原因的反馈，则被评为**不正确（0）**。

所有评价者均为拥有语言学背景的中文母语者，其英语都达到了六级或专业四级水平，并且了解跨语言反馈生成任务。此外，为了确保评估的一致性和可靠性，我们还进行了Kappa 统计分析（0.8172），保证了评价者之间较高的一致性。


(c) 2024 中国中文信息学会计算语言学专业委员会

#### 4.3.3 大语言模型评估方法

由于传统的基于参考的表层相似度指标（BLEU）与人类判断的相关性相对较低，而人类评估的耗时较长且成本较高。因此，对于需要一定创造力和多样性的跨语言反馈评语生成任务，本文尝试使用GPT 3.5-Turbo和GPT 4-Turbo等大型语言模型作为评估的指标，这不仅可以验证我们的方法在更广泛的语境中的有效性，也可以探索如何利用LLM的强大能力来进一步提升跨语言反馈生成的质量。

本文分别采用了无监督评估和有监督评估两种方式：在无监督评估中，我们参考Liu(Liu et al., 2023)和Wang(Wang et al., 2023)提出的大语言模型评估框架，分别探索了在Prompt中指定评估维度和未指定评估维度下大语言模型的评估效果；对于有监督评估，我们将上小节所述的人工评估标准作为Prompt的一部分指导大语言模型进行评分。经过验证发现，使用大语言模型评估与人类评估的一致性较低，因此我们仅在附录C中报告和分析了这部分结果。

## 5 分析

本章使用如4.3节所述的评价指标对模型反馈评语生成效果进行了评估。由于大语言模型方法在与人类教师评语的语言一致性、评语的简洁性、解释的准确性等方面的问题（详见附录B），在本节中我们仅对预训练语言模型方法进行了分析，本文所述预训练语言模型方法的BLEU分数和BERTScore（包括精确度，召回率以及F1分数）评估结果如表3所示。

| | mBART | | | | mT5 | | | |
|---|---|---|---|---|---|---|---|---|
| | BLEU | BERTScore | | | BLEU | BERTScore | | |
| | | P | R | F1 | | P | R | F1 |
| Bare | 47.739 | 84.497 | 83.463 | 83.906 | 40.521 | 83.668 | 82.787 | 83.158 |
| Edit | 43.472 | 85.571 | 84.154 | 84.784 | 42.126 | 84.946 | 84.473 | 84.625 |
| EV | 45.141 | 84.135 | 83.381 | 83.700 | 41.180 | 83.515 | 82.664 | 83.011 |
| GTs | 48.951 | 86.973 | 86.144 | 86.490 | **47.953** | **85.662** | 84.693 | **85.100** |
| GP+Edit | 49.581 | 86.942 | 85.979 | 86.391 | 46.211 | 84.972 | **84.725** | 84.781 |
| GP+EV | 47.546 | 86.793 | 85.153 | 85.896 | 45.901 | 85.116 | 84.329 | 84.644 |
| Edit+EV | **50.317** | **86.982** | **86.308** | **86.589** | 42.693 | 84.198 | 83.413 | 83.736 |
| GP+Edit+EV | 48.503 | 86.803 | 85.961 | 86.313 | 44.538 | 85.069 | 83.961 | 84.442 |

表 3: 预训练语言模型实验结果

通过对mBART模型结果的分析，我们观察到当模型输入中组合附加修正编辑和线索词语（Edit+EV）时，BLEU分数达到最高（50.317），而BERTScore的精准率、召回率和F1分数分别为86.982、86.308和86.589，均高于其他单一策略或组合策略的分数。这一结果表明，修正编辑和线索词语的拼接组合的加入显著提升了模型生成反馈的质量和准确性，这可能是因为附加的语法信息为模型提供了关键的上下文支持。

对于mT5模型而言，采用单独附加语法术语（GTs）的输入策略表现出最优的BLEU分数（47.953），并在BERTScore的精准率和F1分数上获得了相对较高的评分（分别为85.662和85.100）。这一现象说明了语法术语信息对于mT5生成反馈评语有极大的辅助作用。

当采用组合策略（GP+Edit+EV）时，我们注意到尽管这一组合在BLEU分数上均不是效果最佳的情况，但在mBART和mT5模型上的BERTScore F1分数有所提升，分别达到86.313和84.442。这表明了一个多元化的输入组合可能为模型平衡精确度与召回率提供了额外的帮助，尽管这种平衡并未在BLEU分数上反映出显著提升。

在两种模型的对比中，mBART在大多数评价指标上均优于mT5，这可能表明mBART在处理英文到中文的跨语言反馈生成任务上具有更为适宜的架构和更强的处理能力。

鉴于人工评估的时间和成本，我们仅对在mBART和mT5两个模型上取得最佳效果的情况进行了人工评价，其准确率分别为62.84和59.27。

第二十三届中国计算语言学大会论文集，第134页-第149页，太原，中国，2024年7月25日至28日。
卷1：主会论文
(c) 2024 中国中文信息学会计算语言学专业委员会

### 5.1 流水线中的前置模型效果

**修正编辑和线索词语预测模型效果**

由于预测线索词语需要依赖修正编辑，本节将二者合并进行探讨。目前，最先进语法错误纠正模型的准确率约为67.5%。此外，由于在反馈评语数据集中有部分实例是对于语言习惯的评语，即其原始的写法在语法上并不完全错误但是存在更符合习惯的说法，这部分实例也是目前的GEC模型无法提供更多修正信息的。总的来说，GECToR能够标记出语法错误修正的实例不及全部数据条目的50%，如表4所示。

| | 反馈评语数 | 错误修改数 | 错误修改比例 |
|---|---|---|---|
| 训练集 | 4,868 | 2,383 | 48.9523% |
| 开发集 | 170 | 80 | 47.0588% |
| 测试集 | 215 | 90 | 45.5814% |

表 4: GECToR修正错误数量及比例

虽然在全部数据中仅有一部分数据含有附加的修正编辑和线索词语信息，但是我们发现这些仅有部分数据带有的信息仍然起到了重要作用。如表5所示，对于mBART模型，尽管添加修正编辑和线索词语导致了在全部测试数据上效果的下降，但是针对带有附加信息的部分实例，效果的下降大大缓解；对于mT5模型，带有附加信息的实例比整个测试集展现出了更好的效果。因此，可以推测的是随着GEC模型效果的提升，这一方法势必带来更好的跨语言反馈评语生成效果。

| | mBART | | mT5 | |
|---|---|---|---|---|
| | 全部数据 | 标记附加信息部分 | 全部数据 | 标记附加信息部分 |
| Bare | 47.7391 | 48.5225 | 40.5210 | 41.0460 |
| Edit | 43.4723 (-4.2668↓) | 45.2130 (-3.3096↓) | 42.1260 (1.6050↑) | 46.6778 (5.6318↑) |
| EV | 45.1406 (-2.5985↓) | 48.1155 (-0.4070↓) | 41.1802 (0.6593↑) | 43.2662 (2.1803↑) |

表 5: 直接微调和以修正编辑或线索词语作为附加信息的结果分析

**语法术语预测模型效果** 本研究基于附加额外一层线性层的RoBERTa模型，通过修改损失函数的计算方式和索引位置的计算策略，提高了其预测出现频率Top-10语法术语的准确性。最终，语法术语预测的效果如表6所示，该表显示了所提交的模型在开发数据集上的多标签性能。

如表7所示，当使用全部而非Top-10的语法术语作为生成模型输入中的附加信息，生成模型的BLEU值相较于目前的结果有较大幅度的提升，因此，未来对于预测语法术语的多分类任务，提高其准确性或是增加可预测语法术语的数量（如前20个高频率的语法术语）都将对提升反馈评语生成模型的表现提供帮助。

| EMR | Precision | Recall | F1 Score |
|---|---|---|---|
| 18.82 | 85.68 | 72.93 | 78.55 |

表 6: 中文语法术语预测模型结果评估

| | mBART | mT5 |
|---|---|---|
| GTs-ALL | 55.0496 | 51.9758 |

表 7: 使用全部语法术语推理结果

### 5.2 mBART和mT5表现的差异分析

我们观察到，在跨语言反馈评语生成任务中，尽管mT5的模型参数量较大，但其表现却不如参数量较小的mBART。这可能由以下几个原因导致：首先，mBART是专为机器翻译任务设计并进行预训练的，其训练目标与数据处理方式与跨语言反馈生成任务的需求更为吻合。这种专门化的预训练架构可能使得mBART在相关任务上具有更优的表现。相比之下，虽然mT5是一个用途广泛的多语言模型，其预训练任务和目标可能与跨语言反馈生成任务的具体需求不完全匹配。此外，mBART在微调和推理阶段需要指定输入和输出语言，而mT5仅将任务视为通


(c) 2024 中国中文信息学会计算语言学专业委员会

用的文本到文本生成，这可能增加了任务的复杂性。再者，由于mBART与mT5在模型结构上的差异，mBART可能因其结构或内部机制的优化，在处理特定类型的语言生成任务时，能更有效地捕捉语言间的转换和生成规律。

另外一个有趣的现象是，将完全相同=的提示信息附加到模型输入中对mBART和mT5两个模型产生了截然不同的影响。对mBART模型而言，添加修正编辑和线索词语后，其BLEU值仅有很小幅度提升，甚至不及直接对初始模型进行微调的表现效果。相反，mT5模型在加入这些信息后，其BLEU值显著高于直接微调的原始模型。

我们推测，这种差异主要源于模型架构的不同。尽管mT5和mBART均基于Transformer架构的Seq2Seq模型，它们处理输入和输出的方式却有所区别。mT5设计为将所有自然语言处理任务视为文本到文本的转换，利用前缀提示来明确任务类型，因此更适合处理结构化的输入，这使得模型可以将这些输入视为明确的任务指令。相反，mBART以自编码方式预训练，专注于文本恢复，如重构乱序或掩盖的文本，这使其可能更适合对输入进行理解，而对输入的不同表示形式敏感度较低。

此外，mT5设计初衷就是处理多样化文本输入，并产生相应的输出。附加信息丰富了输入文本的上下文，使mT5能够更有效地利用这些信息，从而适应多变的输入格式。相对而言，mBART可能更依赖输入的一致性和结构，它在处理更直接和传统的文本生成任务中可能表现更佳。然而，在生成反馈评语的任务中，不是所有输入都包含修正编辑和关键词，这种结构上的不一致可能影响了mBART对输入的理解，从而使得额外信息反而干扰了对原始文本的处理，导致性能下降。

对于适用于文本到文本转化的mT5模型而言，输入中附加语法术语有很大概率在输出中被直接使用，因此其取得了最佳的性能。而对于在理解输入方面能力更强的mBART模型，输入中附加的修正编辑和线索词语可能为其提供了更丰富的上下文逻辑信息，使其表现更好。

### 5.3 更多附加信息导致模型效果下降

根据我们的假设，当模型的输入包含更多附加的提示信息时，其生成效果理应有所提升，这也是很符合直觉的。但是，我们发现在输入中添加多种附加信息可能会导致模型效果不及添加其中单一附加信息更有效，即添加更多附加信息反而导致了模型效果的下降。例如，表3中mBERT模型附加全部语法术语、修正编辑和线索词语的情况不及只附加修正编辑和线索词语的情况，这一现象在mT5模型中表现得更加明显，附加多个信息的策略均不及仅单独使用语法术语作为附加信息的效果。

本研究认为导致这一现象的是由于信息过载导致的注意力分散，即将语法术语、修正答案和线索词语同时附加到输入可能导致信息量过大。对于基于Transformer架构的模型来说，输入的每个组成部分都会通过注意力机制相互影响。如果输入信息过于复杂或不直接相关，模型可能难以从中提取有效的特征，从而影响学习效果。虽然单独使用三者对模型生成反馈评语都是有用的信息，但它们可能在语义上存在重叠或者相互干扰，这会使模型在处理输入时产生混淆，难以区分哪部分信息更重要。如果语法术语、修正答案和线索词语被同时附加，模型的注意力可能会在这三类信息之间分散，使得关键信息的权重降低。

### 5.4 案例分析

表8展示了基于流水线的预训练语言模型在生成反馈评语时，使用不同附加信息，导致生成效果差异的一个典型的示例。

该表前两行分别为学习者语句和人类教师提供的反馈评语，之后五行分别为不同方式下mBART模型的附加信息及生成的反馈评语等。表中红色字体为评语中存在错误的内容，而绿色字体为正确且其他评语未涉及的内容。

在学习者语句中，“which should be for learning”是修饰宾语“time”的定语从句，而“work”应该作为目的状语，解释他们花时间的目的。但是，学习者错误的使用了介词“to”与“learning”进行搭配。不含任何附加信息的端到端方法生成的解释完全错误；包含语法术语或修正编辑的方法正确的解释了修改方式但错误原因分析错误；而附加线索词语的方式首次正确识别出关联错误的词语为“spend”；取得最佳表现的修正编辑+线索词语的方式进一步完善了修改方式，即提示出可以改为working或on working。这一结果与添加各个附加信息的初衷相符，证明了方法的有效性。


(c) 2024 中国中文信息学会计算语言学专业委员会

| | | | |
|---|---|---|---|
| They spend time which should be for learning [to work], thus leading to decline of grades. | | | |
| 使用结构<介词+动名词>或<动名词>代替<to不定式>来表达“在某事上浪费时间”。在字典中查找<动词><<spend>>的正确用法。 | | | |
| **Bare** | / | 由于<动词><<learn>>是一个<及物动词>，因此<宾语>前面不需要有<介词>。在字典中查找<动词><<learn>>的用法。 | × |
| **GT** | <to不定式>、<动名词> | 使用<to不定式>的<动名词>形式来表达“为了某事努力工作”。 | ✓ |
| **Edit** | to work → working | 使用结构<介词+动名词>或<动名词>代替<to不定式>来表达学习的东西。在字典中查找<动词><<work>>的用法。 | ✓ |
| **EV** | spend、for | 使用结构<动名词>代替<to不定式>来表达“在某事上浪费时间”。在字典中查找<动词><<spend>>的正确用法。 | ✓ |
| **Edit+EV** | to work → working spend、for | 使用结构<介词+动名词>或<动名词>代替<to不定式>来表达“在某事上浪费时间”。在字典中查找<动词><<spend>>的正确用法。 | ✓ |

表 8: 案例分析

## 6 结语

在本研究中，我们提出了一种新颖且具有挑战性的跨语言反馈评语生成任务（TLFCG）。首先，我们基于GPT-4模型，利用人工标注的语法术语，对现有资源进行了翻译，并通过人工校对提升翻译质量，从而构建了首个英-中跨语言反馈评语数据集。为应对这一任务，我们使用了两种多语言预训练模型，并探讨了不同附加信息对模型效果的影响。通过广泛的实验验证，我们的方法能够有效地处理跨语言反馈评语生成任务。通过对结果的深入分析，我们期望为自然语言处理社区贡献更多洞见。未来，我们计划探索输入中多个提示信息的拼接与排序对模型的影响，并进一步分析本文所提方法对不同错误类型生成反馈评语的效果差异。

## 致谢

感谢各位评审专家对本文的细致审阅和宝贵建议。他们的建议丰富了本文的内容，提高了本文的质量，更加深了我们的研究视角。衷心感谢他们对本研究的贡献。

## 参考文献

Nikolay Babakov, Maria Lysyuk, Alexander Shvets, Lilya Kazakova, and Alexander Panchenko. 2022. Error syntax aware augmentation of feedback comment generation dataset, December.

Shabnam Behzad, Amir Zeldes, and Nathan Schneider. 2022. Sentence-level Feedback Generation for English Language Learners: Does Data Augmentation Help?, December.

Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. Automatic Annotation and Evaluation of Error Types for Grammatical Error Correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 793–805, Vancouver, Canada, July. Association for Computational Linguistics.

Yuejiao Fei, Leyang Cui, Sen Yang, Wai Lam, Zhenzhong Lan, and Shuming Shi. 2023. Enhancing Grammatical Error Correction Systems with Explanations. May.

Kazuaki Hanawa, Ryo Nagata, and Kentaro Inui. 2021. Exploring Methods for Generating Feedback Comments for Writing Learning. In *Proceedings of the 2021 Conference on Empirical Methods*


(c) 2024 中国中文信息学会计算语言学专业委员会

*in Natural Language Processing*, pages 9719–9730, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Mana Ihori, Hiroshi Sato, Tomohiro Tanaka, and Ryo Masumura. 2023. Retrieval, Masking, and Generation: Feedback Comment Generation using Masked Comment Examples. In *International Conference on Natural Language Generation*.

S. Ishikawa. 2013. The ICNALE and Sophisticated Contrastive Interlanguage Analysis of Asian Learners of English. March.

Kunitaka Jimichi, Kotaro Funakoshi, and Manabu Okumura. 2023. Feedback comment generation using predicted grammatical terms. In *International Conference on Natural Language Generation*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach, July.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. GPTEval: NLG Evaluation using GPT-4 with Better Human Alignment. March.

Ryo Nagata, Kentaro Inui, and Shin'ichiro Ishikawa. 2020. Creating Corpora for Research in Feedback Comment Generation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 340–345, Marseille, France, May. European Language Resources Association.

Ryo Nagata, Masato Hagiwara, Kazuaki Hanawa, Masato Mita, Artem Chernodub, and Olena Nahorna. 2021. Shared Task on Feedback Comment Generation for Language Learners. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 320–324, Aberdeen, Scotland, UK, August. Association for Computational Linguistics.

Ryo Nagata, Masato Hagiwara, Kazuaki Hanawa, and Masato Mita. 2023. A Report on FCG GenChal 2022: Shared Task on Feedback Comment Generation for Language Learners. In *International Conference on Natural Language Generation*.

Ryo Nagata. 2019. Toward a Task of Feedback Comment Generation for Writing Learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3206–3215, Hong Kong, China, November. Association for Computational Linguistics.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. 2020. GECToR – Grammatical Error Correction: Tag, Not Rewrite, May.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A Method for Automatic Evaluation of Machine Translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual Translation with Extensible Multilingual Pretraining and Finetuning, August.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is ChatGPT a Good NLG Evaluator? A Preliminary Study.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer, March.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT, February.

CCL 2024

## 附录A.大语言模型反馈评语生成

为了探索大语言模型在跨语言反馈评语生成任务上的表现，我们选择使用OpenAI的闭源商用大语言模型GPT-3.5-turbo和GPT-4-turbo进行实验，二者均采用作者撰写本文时的最新版本，即gpt-3.5-turbo-0125和gpt-4-turbo-2024-04-09。

**大语言模型反馈评语生成Prompt：**

你是一位经验丰富的中国英语教师。你需要对于给定的一个学习者语句和其中指定的错误提供具体且简洁的中文反馈评语，让学习者能够理解这个错误的本质原因。学习者语句中使用方括号[和]标记错误位置。
每一个反馈评语应该包含：为什么学习者语句中指定的结构是错误或不合适的，其本质的原因或者规则是什么?
请直接给出反馈评语。
如下是几个例子：
学习者语句：And how to propaganda and let people [agree this] rule?
反馈评语：由于<动词><<agree>>是一个<不及物动词>，所以<介词>需要在<宾语>之前。在字典中查找<动词><<agree>>以找到合适的<介词>。
学习者语句：We must consider other people especially [at] the public places.
反馈评语：<介词><<at>>可以与<<place>>一起使用来指示某事发生的地点，但更常见的是使用'in' 代替。
学习者语句：It was miserable and I thought that I did n't want to face [to] such a situation.
反馈评语：由于<动词><<face>>是一个<及物动词>，因此<宾语>不需要<介词>。
学习者语句：{{学习者语句}}
反馈评语：

在上述指令中，我们将需要模型生成反馈评语的学习者语句以示例同样的形式输入模型，有助于其仅生成反馈评语而不包含其他内容。为了使大语言模型得到更好的效果，我们同样在System Prompt中添加了ChatGPT的指令：

You are ChatGPT, a large language model trained by OpenAI, based on the GPT-4 architecture.\n Knowledge cutoff: 2023-12\n Current date: 2024-05-12

## 附录B.大语言模型方法结果与分析

由于以下原因我们决定仅使用大语言模型辅助人工进行评语翻译，而最终使用预训练语言模型进行反馈评语生成：

**与人类教师评语的语言一致性**：面向语言学习者的反馈评语生成不仅要准确，还要具备教育意义和可理解性，以确保学习者能从中获得最大的学习效益。我们的研究显示，经过任务数据微调的预训练语言模型能够更好地模仿人类教师使用的自然和教育性语言风格。

在实验中，我们人工比较了预训练模型生成的评语与真实教师评语的语言风格和表达方式。结果表明，预训练模型的输出在语言与人类教师的评语更为接近，这有助于增强反馈的接受度和效果。相反，大语言模型虽然能够生成语法结构正确的句子，但其风格和用词往往缺乏针对性和指导性细节，有时可能显得过于机械或与教学语境不够贴合。虽然大语言模型在通用领域的能力远高于预训练语言模型，但由于对其的微调需要大量的计算资源和训练数据。因此，导致目前的情况下难以对其进行微调，以使其适配到跨语言反馈评语生成任务中，更加难以对齐人类教师反馈评语的内容、语言和组织形式。

**评语的简洁性**：在实际语言教学场景中，反馈评语的长度直接影响了学习者的阅读兴趣和理解程度。一些研究已经表明，大语言模型在表达相同的意思时倾向于生成更加长的内容。尽管我们已经在Prompt中要求大语言模型生成尽可能简洁的评语并添加了一些示例，但是其生成的评语长度仍然达到了人工标注和预训练语言模型的近两倍，而这会大大影响学习者的学习体验和效果。预训练语言模型生成的评语长度则于人工标注结果更为接近，结果如表9所示。

**解释的准确性**：在生成具体针对学习者错误的反馈评语时，准确性至关重要。我们评估了在few-shot情况下，大语言模型GPT-3.5、GPT-4和使用修正编辑+线索词语（Edit+EV）流水线的预训练语言模型mBART在本任务上的表现。在这里我们额外添加了宽松的人工评价指

| 模型/方法 | 平均评语长度（字符） |
|---|---|
| 人工标注 | 59.72 |
| mBart-Large（Edit+EV） | 56.26 |
| GPT-3.5 （few-shot） | 115.09 |
| GPT-4 （few-shot） | 119.45 |

表 9: 不同方法生成反馈评语的平均长度

标，即生成的反馈评语中只要包含有错误相关的信息，我们即认为它是正确的。我们的实验表明，GPT-3.5表现较差，而GPT-4展现出显著的鲜果提升，特别是在严格的人工评估中超过了GPT-3.5约32.56%。同时，使用Edit+EV方法的mBART模型也表现出色，特别是在宽松评估中接近80的高分。这一结果突显了使用先进的预训练模型和专门的优化策略对于提高反馈准确性的重要性。

| 模型/方法 | 人工评估（严格） | 人工评估（宽松） |
|---|---|---|
| GPT-3.5 （few-shot） | 37.21 (25.42↓) | 56.98 (23.00↓) |
| mBart-Large（Edit+EV） | 62.84 | 79.98 |
| GPT-4 （few-shot） | 69.77 (7.14↑) | 89.53 (9.55↑) |

表 10: 大语言模型于预训练语言模型效果的人工评估

从上表可以看出，GPT-4在few-shot设置下对反馈生成任务的适应性和效果显著优于GPT-3.5，而mBART-Large则利用其特有的附加修正编辑和线索词语的策略，表现出稳定而高效的反馈生成能力。在跨语言反馈评语生成任务中，仅拥有0.68B参数的预训练语言模型mBART在经过微调后，效果较大幅度超过拥有175B参数量GPT-3.5，较小幅度低于包含约1800B参数规模的GPT-4，我们认为这足够令人兴奋。这意味着相较于通用的大语言模型，预训练模型在本研究的特定任务上进行了深入优化，同样能够很好地适应语言学习反馈的需求。虽然我们的结果尚不及最新的GPT-4-Turbo模型，但其仍然提供了有价值的见解和创新，我们希望以此为未来的研究奠定基础。

**其他原因**：

1.资源的可持续性：大型模型如GPT-3.5和GPT-4在运行时需要大量的计算资源，这可能不适合所有使用场景，特别是资源有限的环境。相比之下，预训练语言模型通常具有更低的资源需求，更适合长期可持续发展。

2.可定制性：预训练模型更容易根据特定的教育需求进行调整和定制。例如，可以根据学习者的具体错误类型或学习阶段调整模型参数，更好地适应教育场景。

随着大语言模型的发展和微调技术研究，大模型在跨语言反馈评语生成任务上的表现会更加优异。我们也将持续关注相关的研究成果，并将其作为我们未来的研究工作。

## 附录C.大语言模型评估方法

### C.1 Prompt内容

**大语言模型有监督评估Prompt**：

> 你是一位具有语言学背景的评价者。你需要对反馈评语生成模型针对语言学习者语句给出的反馈评语进行评价。
> 反馈评语生成任务是对于给定的一个学习者语句和其中指定的错误提供具体且简洁的中文反馈评语，让学习者能够理解这个错误的本质原因。学习者语句中使用方括号[和]标记错误位置。
> 每一个反馈评语应该包含：为什么学习者语句中指定的结构是错误或不合适的，其本质的原因或者规则是什么?
> 评价使用0至2的评分系统基于特定评估标准对每条生成的反馈评语进行评估：完全正确


(c) 2024 中国中文信息学会计算语言学专业委员会

（2）、部分正确（1）或不正确（0）。其中，完全正确（2）指反馈评语不仅包含了与参考内容相似的信息，而且未包括任何与错误无关的内容。即便反馈中包含了参考内容未提及的相关信息，只要这些信息直接关联到错误，也被认定为完全正确。部分正确（1）的评分意味着反馈评语基本准确，但需要简单编辑以提高准确度；例如，如果反馈正确指出了句子的错误，且只需修改几个词便能让内容更准确，则归于此类。而对于那些与参考内容完全无关，未能指出错误原因的反馈，则被评为不正确（0）。
请直接给出评分，而不包含其他内容。
学习者语句："{{学习者语句}}"
参考反馈评语："{{参考反馈评语}}"
待评反馈评语："{{待评反馈评语}}"
评分（0-2）：

**大语言模型无监督评估且不提供评估维度Prompt：**

你是一位具有语言学背景的评价者。你的任务是根据指标对反馈评语生成模型针对语言学习者语句给出的反馈评语进行评价。
反馈评语生成任务是对于给定的一个学习者语句和其中指定的错误提供具体且简洁的中文反馈评语，让学习者能够理解这个错误的本质原因。学习者语句中使用方括号[和]标记错误位置。
请确保您仔细阅读并理解这些说明。请在审阅时保持本文档处于打开状态，并根据需要进行参考。
评估步骤：
1.仔细阅读学习者语句，找出错误位置错误的原因。
2.阅读反馈评语并将其与学习者语句中的错误进行对照。检查反馈评语是否解释了学习者语句中的错误和原因，以及是否以清晰且合乎逻辑的方式呈现。
3.根据评估维度，按照0到2的范围打分，其中0为最低，2为最高。
学习者语句："{{学习者语句}}"
反馈评语："{{待评反馈评语}}"
评估表（仅给出综合分数而不需要任何说明）：
- 综合分数（0-2）：

**大语言模型无监督评估且提供评估维度Prompt：**

你是一位具有语言学背景的评价者。你的任务是根据指标对反馈评语生成模型针对语言学习者语句给出的反馈评语进行评价。
反馈评语生成任务是对于给定的一个学习者语句和其中指定的错误提供具体且简洁的中文反馈评语，让学习者能够理解这个错误的本质原因。学习者语句中使用方括号[和]标记错误位置。
请确保您仔细阅读并理解这些说明。请在审阅时保持本文档处于打开状态，并根据需要进行参考。
评估维度：
1.精确性（0-2），即评价反馈评语是否正确指出了学习者语句中的错误，并且正确解释了为什么这个结构是错误的或不合适的，前提是解释中的错误原因是真实存在的。评分标准：2-完全正确，清楚地指出错误并提供了正确的规则或原因；1-部分正确，可能识别了错误，但解释不够详细或部分错误；0-错误识别或解释错误，误导学习者。
2.明确性（0-2），即评价反馈评语是否表达清晰，无歧义，学习者能够轻易理解。评分标准：2-非常清晰，使用简洁的语言明确表达错误原因；1-表达较为清晰，但可能存在少量模糊或复杂的表述；0-表达含糊，难以理解，可能导致学习者混淆。
3.相关性（0-2），即评价反馈是否专注于学习者语句中标记的错误，没有偏离主题。评分标准：2-完全相关，专注于指定错误，没有无关内容；1-大体相关，但包含一些不必要或边缘的信息；0-反馈中包含大量无关信息，与标记的错误无关。
4.教育价值（0-2），即评价反馈是否提供了帮助学习者改正错误和避免未来错误的具体建

CTI 2024

议或例子。评分标准：2-不仅指出错误，还提供了如何改正的建议或规则的深入解释；1-提供了基本的改正建议，但缺乏深入解释；0-没有提供改正建议或解释，无助于学习者的长期学习。
评估步骤:
1.仔细阅读学习者语句，找出错误位置错误的原因。
2.阅读反馈评语并将其与学习者语句中的错误进行对照。检查反馈评语是否解释了学习者语句中的错误和原因，以及是否以清晰且合乎逻辑的方式呈现。
3.根据评估维度，按照0到2的范围打分，计算其平均值并取整（综合分数），其中0为最低，2为最高。
学习者语句："{{学习者语句}}"
反馈评语："{{待评反馈评语}}"
评估表（仅给出综合分数而不需要任何说明）：
- 综合分数（0-2）：

### C.2 结果与分析

使用大语言模型（GPT 3.5和GPT 4）分别对预训练语言模型和大语言模型生成的反馈评语进行评估的结果如下表所示。

| | | 评估模型 | | | | | |
|---|---|---|---|---|---|---|---|
| | | GPT-3.5 | | | GPT-4 | | |
| | | 无监督 | | 有监督 | 无监督 | | 有监督 |
| | | 无维度 | 含维度 | | 无维度 | 含维度 | |
| 生成模型 | GPT-3.5（few-shot） | 94.76 | 60.47 | 69.70 | 74.42 | 59.30 | 46.18 |
| | mBart-Large（Edit+EV） | 77.91 | 58.14 | 65.59 | 43.02 | 66.28 | 70.24 |
| | GPT-4（few-shot） | 98.83 | 84.88 | 60.12 | 97.67 | 88.37 | 85.88 |

表 11: 大语言模型评估结果

在无监督评估过程中，当Prompt没有指明评估维度时，GPT-3.5模型和GPT-4模型都对大语言模型生成的反馈评语给出了较高的分数。这是由于在没有评分参考和示例时，大语言模型并没有清晰地理解反馈评语生成任务，因此更加倾向于大语言模型自身生成的答案。但当Prompt包含评估维度时，GPT-4模型降低了对大预言模型结果的分数，并提高了对预训练语言模型的打分，这说明评分维度的加入使得大语言模型更加明确了这一任务及其评估方式。

在有监督的评估过程中，通过以人工标注的反馈评语作为参考，大模型可以更加了解任务的模式，以更加客观和符合任务要求的方式进行评估。一个有趣的现象是大语言模型都更加倾向于自己生成反馈评语，这说明大语言模型在作为模型效果的评估者时可能会带有偏见。

总体而言，应该使用能力更强的模型对能力较弱的模型进行评估，正如老师评价学生一样，所以在此我们倾向于采信GPT-4作为评估者的结果，尽管其会更偏向自己生成的答案。GPT-4与人工评估的Kappa一致性分别为4.52、15.43和34.30，表明大语言模型和人类评估者之间的一致性是公平的。这也证明通过增加评估指标和参考示例可以提高大模型对评估任务的理解和与人类评估的一致性。


(c) 2024 中国中文信息学会计算语言学专业委员会

# 文本样式和主题框架引导下的大模型辅助儿童新闻生成

**杜晓蒙[1]，于东[1,*]，刘鹏远[1,2]**
1.北京语言大学信息科学学院，北京，100083
2.国家语言资源监测与研究平面媒体中心，北京，100083
202221198701@stu.blcu.edu.cn, yudong@blcu.edu.cn, liupengyuan@blcu.edu.cn

## 摘要

主流新闻内容多针对成年人设计，不易于儿童理解，难以满足其阅读需求。对此，我们提出了一种基于主题的儿童新闻篇章结构框架（TNC-LLM）。该框架融合了文本样式定义（TSD）和主题类别定义（TCD）两大核心模块，TSD模块采用多种机器学习算法，从不同粒度分析文本样式风格和段落布局等特点，TCD模块针对不同主题进行了内容分析，以揭示儿童新闻的写作特点和内容的倾向性，确保内容的教育性和适宜性。本文实验主要评估了ChatGPT3.5等四个模型在将成年人新闻转换为面向儿童的新闻的性能。实验结果表明，TNC-LLM在儿童新闻内容生成任务中对内容的准确性、文本的趣味性以及教育性等关键维度有显著提升。此外，该框架具有普适性，能够应用于不同类型的大型语言模型。

**关键词：** 文本生成；大型语言模型；儿童新闻

# Text Styles and Thematic Framework Guided Large Modeling to Aid Children's News Generation

**Xiaomeng Du[1], Dong Yu[1,*], Pengyuan Liu[1,2]**
1.Faculty of Computer Science,Beijing Language and Culture University,Beijing,100083
2.National Language Resources Monitoring and Research Center for Print Media,Beijing,100083
202221198701@stu.blcu.edu.cn, yudong@blcu.edu.cn, liupengyuan@blcu.edu.cn

## Abstract

Mainstream news content is mostly designed for adults, which is not easy for children to understand and difficult to meet their reading needs. In this regard, we propose a topic-based children's news chapter structure framework (TNC-LLM). The framework integrates two core modules, Text Style Definition (TSD) and Theme Category Definition (TCD). The TSD module adopts various machine learning algorithms to analyze the characteristics of text style style and paragraph layout from different granularities, and the TCD module conducts content analysis for different topics to reveal the writing characteristics and content tendency of children's news and to ensure that the

*为通讯作者
**基金项目:**2023年国际中文教育研究课题重点项目（23YH38B）；北京语言大学梧桐创新平台(中央高校基本科研业务费)(21PT04)
©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

content is educational and appropriate. The experiments in this paper mainly evaluate the performance of four models, including ChatGPT3.5, in converting adult news into child-oriented news. The experimental results show that TNC-LLM significantly improves the key dimensions of content accuracy, text interest, and educational appropriateness in the children's news content generation task. In addition, the framework is universal and can be applied to different types of large-scale language models.

**Keywords:** Text Generation , Large Language Models , Children's News

## 1 引言

阅读新闻是儿童提高社会能力和阅读能力的重要途径。皮亚杰(1981)研究表明儿童读物应与儿童所处的认知发展阶段相匹配，儿童读物应根据儿童的年龄、认知水平和兴趣进行个性化设计(Pelletier, 2004; 饶芳芳, 2023)。现有的新闻报道大多以成年人为主要受众，其形式结构上不满足儿童阅读需求，难度上与儿童阅读能力不适配。如图1所示，成年人版新闻和儿童版的新闻在语言风格、段落布局等有着明显的差异性。如何利用技术手段实现图1的成年人新闻到儿童新闻转写，成为具有应用价值的研究课题。此外，目前面向儿童的新闻文本数量有限，无法满足儿童阅读新闻的需求。

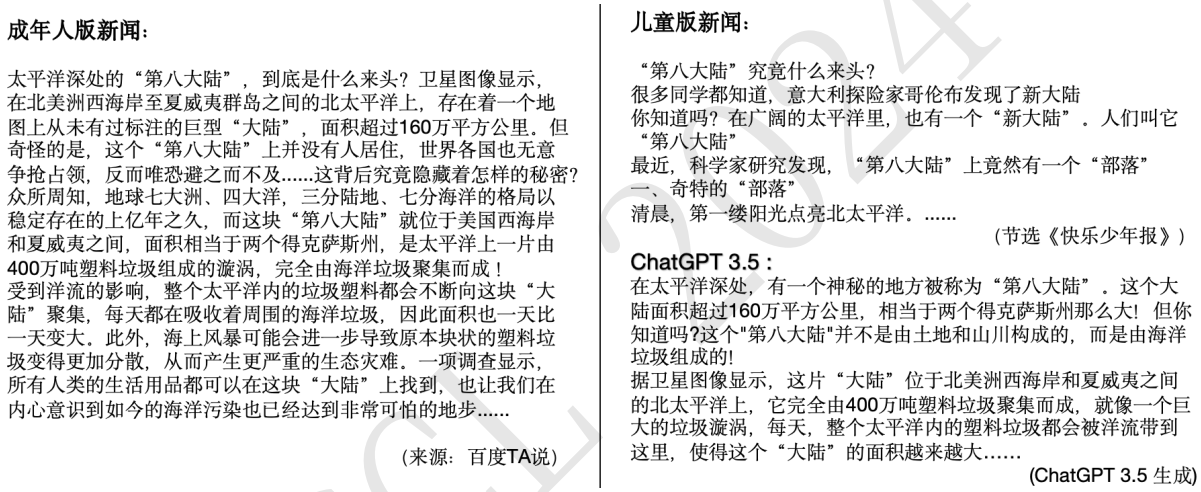



图 1. 同一主题新闻的成年人版和对应儿童版示例

目前的研究领域中，针对将文本内容转化为适合儿童阅读理解的版本，主要采用了文本简化的方法。通过将原文中的复杂词汇替换为更简单、更易于儿童理解的词语，来降低文本的阅读难度(Xu et al., 2018)。在长文本简化任务上，采用将长篇文本内容压缩成简短、精炼的摘要的方法来降低文本难度(Zhong et al., 2020)。尽管这些方法在提高文本可读性方面取得了一定的成效，但它们在内容的敏感性、教育价值、语言表达方式、以及段落排版布局等方面，常常忽视了对儿童认知发展特性和兴趣偏好的系统性考虑和深入分析。

大型语言模型（LLM）通过在海量文本数据上的预训练，展现了卓越的语义理解和生成能力，这使得它们在文本生成和文本创作方面表现出色(Zhao et al., 2023)。无需特定任务训练，可以有效的解决小模型在文本生成所面临的问题如上下文连贯性等。在新闻文本生成方面，大型语言模型已经被应用于自动化地生成新闻报道。例如，一些新闻机构利用LLM来撰写财经报告、体育赛事结果等结构化新闻(Newman, 2024)。此外，LLM也被用于辅助记者进行研究和内容创作，通过生成新闻摘要或提出报道角度来提高工作效率。这些应用为如何利用LLM生成适合儿童阅读的新闻内容提供了有价值的启示。如图1所示，虽然可以通过简单的提示让ChatGPT 3.5基于成年人新闻内容生成适合儿童阅读的版本，但这种方法产生的文本往往只是对原文的简化，并没有充分考虑到儿童新闻在语言表达、风格以及段落结构等方面的特殊需求。


(c) 2024 中国中文信息学会计算语言学专业委员会

针对以上问题，我们提出了一种儿童新闻文本样式定义（TSD）模块。该模块专注于捕捉适合儿童阅读的新闻文本特征，这些特征覆盖了字词选择、句子结构、段落组织等多个层面，TSD模块能够协助LLM基于成年人新闻内容，生成难度适宜、适合儿童年龄的新闻文本。该模块还能优化文本的段落布局，使其更加符合儿童的阅读习惯。此外，本研究还针对不同类别的儿童新闻内容进行了深入分析，提出了主题类别定义(TCD)模块，旨在探讨各种类别下儿童新闻的写作特色。通过细致地剖析每个类别特点为儿童新闻创作提供更为精准的指导和参考。通过融合TSD和TCD两个模块的优势，我们借助LLM在语义理解和长文本生成方面的先进能力，成功地将成年人新闻文本转换成既有趣又具有教育意义，同时适合儿童阅读的新闻文本。我们基于此构建了儿童新闻文本的数据集，并构造了与儿童新闻主题相对应的成年人版的新闻文本。

本文的贡献：第一，提出了基于主题的儿童新闻篇章结构(TNC-LLM)框架，这是大型语言模型在儿童新闻文本生成领域的首次深入应用，经过大量的实验证明，TNC-LLM在文本内容以及趣味性、教育性等多方面得到了显著的提升。第二，设计了文本样式定义(TSD)模块，该模块集合了多种机器学习的算法，可以在字、词和句级别熟练的总结文本样式和难度特征，在段落级别可以准确地总结出文本的布局和格式特点。该模块不仅提升了LLM在文本内容的保留度，还可以将难度适配到小学生阅读水平。第三，设计了主题类别定义（TCD）模块，旨在深入剖析各类新闻内容的特点和撰写风格。该模块能够精准捕捉儿童新闻报道的倾向性、语言风格和内容展现形式，进而基于成年人文本进行动态结构转换，生成既符合儿童兴趣爱好又积极正面的儿童新闻文本，同时传递了有益的主题和价值观。

## 2 相关工作

可控文本生成（CTG）任务是指生成符合语法规则和语义需求，且满足给定约束的自然语言文本，具有重要应用价值。可控文本生成领域在情感控制、主题导向和写作风格调整等关键领域已经取得了显著进展。在风格转换领域(Xu et al., 2018)开创性地提出了一种循环强化学习模型，有效解决了文本风格转换中缺乏成对样本的情感转换难题。(Li et al., 2018)聚焦于风格转换过程中的文本属性，通过替换这些关键词汇可以直接改变文本的性质属性。尽管上述两项研究在风格转换方面取得了成就，但它们主要适用于简单句式结构的转换，且生成内容较为单一。随着需求的增加，研究者寻求更多样化和可控的文本生成技术。(Qian et al., 2022)研究提出了一种使用GPT-2模型的可控文本生成方法，通过属性向量引导生成过程，实现了细致的控制，提高了生成控制的精度。(Dathathri et al., 2020)中采用了基于预训练模型精调的可控文本生成方法，通过在预训练模型上添加属性判别模型，实现了在不重新训练整个模型的情况下对文本新属性的控制。然而，在新闻文本生成上仅仅依赖可控文本生成直接输出的内容，并不能确保其事实准确性。

故事生成任务是人工智能和自然语言处理领域的一个关键研究方向，旨在赋予计算机自动创作出既连贯又引人入胜的故事的能力。(Lebowitz, 1985)是故事生成领域的开创性工作之一，他提出了一种基于规划和学习的故事生成方法，将故事生成过程视为一个规划问题。随后，(Fan et al., 2018)提出了一种创新的故事生成框架，通过整合分层生成、模型融合和自注意力机制，显著提升了故事生成的质量。(Liu et al., 2020)研究中提出了一个以角色为中心的神经网络模型，通过明确捕捉人物信息以及情节与人物之间的关系，增强了故事生成的可解释性和一致性。尽管GPT-2等大型预训练语言模型在文本生成上表现出色，但它们在可控性上仍有不足。在需要知识驱动的生成任务，如故事创作时，这些模型可能会生成逻辑或语法错误的内容。为此(Xu et al., 2020)提出了一种新的框架，该框架允许将外部知识动态地融入语言模型，实现对文本生成过程的精细控制。为了应对现有语言模型在生成长篇故事时的挑战，(Park et al., 2023)提出了一个新的任务——长度可控的故事生成。该任务通过递归地生成故事片段，同时确保整体的连贯性和完整性。通过定量分析，Park证明了其方法LongStory在生成较长故事方面的优势，尤其是在提升连贯性、完整性、相关性以及减少重复性等方面。

经过大规模语料库的预训练，大型语言模型具备了处理各类任务的潜力(Wei et al., 2022)。然而，这种潜力在执行特定任务时可能并不总是明显。因此，设计恰当的任务指令或特定的上下文策略来激发这些潜能变得极为关键。(Chen et al., 2023)的研究表明，提示工程在释放LLM的功能方面发挥着核心作用。在(Kaddour et al., 2023)的研究进一步证实了提示在模型输入中的重要性，精心设计的提示能够显著改变输出结果。(Lu et al., 2022)的研究中揭示了提


(c) 2024 中国中文信息学会计算语言学专业委员会

示中包含的短语和示例序列对模型行为的重大影响。指令调优技术赋予了大型语言模型在缺乏显式样本的情况下执行新任务的能力，通过深化模型对指令的理解，该技术确保了LLMs能够更加高效地适应并完成各种任务。这种方法为我们在儿童文本生成等特定领域的应用提供了新的视角和解决方案。

## 3 数据集构建

本文选择了两份面向儿童读者的新闻出版物——《快乐少年报》和《阳光少年报》作为主要的数据来源。两份报刊风格相似，内容丰富多样，既适合儿童阅读，也符合我们构建儿童新闻数据集的需求。研究涵盖了《阳光少年报》2021年10月至2023年10月的“榜样”、“中国”、“世界”、“文化”和“科学”板块，以及《快乐少年报》2023年5月至12月的“世界大事”、“世界趣事”、“中国大事”、“中国趣事”和“文化”板块。鉴于两类报刊的新闻分类标准不一，我们重新分类了文本，建立了包括“榜样”、“政治”、“经济”、“文化”、“科学”、“社会”、“体育”和“环境保护”在内的八个类别。为确保分类的准确性，我们首先使用GPT-4模型进行新闻文本分类。我们给定了每个类别的定义和特征，如环境保护：着重呼吁儿童关注环境保护问题，如节约能源、减少垃圾、保护动植物等，以及简单易懂的环保知识和行动建议。随后，我们通过人工审核模型的分类结果进行复核和调整，以确定每篇新闻的类别。此外，根据新闻文本在报刊中的排版长度，我们将文本划分为“儿童长新闻”和“儿童短新闻”。如图4给出了详细的数据集统计结果。目前市面上针对儿童的新闻报刊资料仅限于纸质形式，并且是周期性出版。我们制定了长期的收集计划来扩展我们的儿童新闻语料库。如图3为儿童长新闻数量饼图。

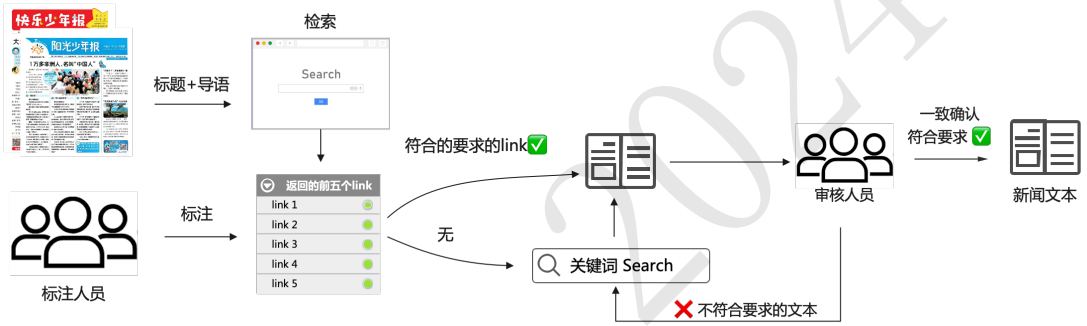



图 2. 成年人版新闻文本标注流程图

在选择成年人新闻文本的过程中，我们采取了自动化检索结合人工审核的方法。如图2为成年人新闻标注流程图。首先提取出儿童新闻标题和摘要作为百度搜索引擎的输入，自动检索并统一返回检索结果的前五个搜索结果，然后由五名硕士生组成的标注团队对返回的结果进行筛选，挑选出与儿童主题相关、内容适宜的新闻，并优先选择知名媒体的新闻。若返回的检索结果中没有符合要求的文本，标注人员根据关键词进行人工检索，找到符合要求的成年人新闻。在初步筛选后，两位语言学背景的硕士生对标注数据进行审核，剔除不合格的文本，并要求标注团队重新标注。通过多轮迭代审核，我们确保了最终结果的准确性和质量。最终，我们收集到了一批高质量、与儿童新闻内容相匹配的成年人新闻文本。所收集的数据数量如图4所示。

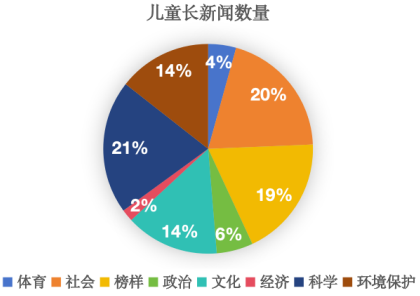



图 3. 儿童长新闻数量饼图

| 文本类别 | 儿童长新闻数量 | 儿童短新闻数量 | 成年人新闻数量 |
|---|---|---|---|
| 体育 | 23 | 29 | 46 |
| 社会 | 107 | 76 | 178 |
| 榜样 | 100 | 11 | 111 |
| 政治 | 30 | 9 | 39 |
| 文化 | 77 | 38 | 115 |
| 经济 | 10 | 4 | 14 |
| 科学 | 110 | 77 | 187 |
| 环境保护 | 77 | 46 | 123 |

图 4. 新闻数据集的统计结果


(c) 2024 中国中文信息学会计算语言学专业委员会

# 4 方法

## 4.1 任务定义和整体框架

假设现在有一组数据集$\{D_i\}_{i=1}^{K}$，每个数据对中包含一篇儿童新闻文本和对应的成年人版新闻报道。我们将$D_i$语料库分为两个部分：$D_{i1}$和$D_{i2}$，其中$D_{i1}$中的文本都共享特有的的样式$\{s_i\}_{i=1}^{K}$，以及在不同主题类别上的特征样式$\{c_i\}_{i=1}^{K}$。任务是根据一篇新闻文本$x$的内容生成为同时具有目标样式s和c的新文本$x'$。

本文中，我们提出的基于主题的儿童新闻篇章结构框架（TNC-LLM）主要分为三个阶段。我们的语料库$D_i$包含两个部分：$D_{i1}$和$D_{i2}$，其中$D_{i1}$是儿童新闻文本语料，$D_{i2}$是对应成年人版新闻文本语料。第一阶段我们提出了文本样式定义(TSD)模块，该模块在语料$D_{i1}$的基础上，分别从字、词、句和段落四个层面综合考虑了文本样式的风格特征。由于文本样式定义部分的文本统计是比较随机和广泛的，所以我们将总结出来的文体特征作为文章的文本样式的风格特征。第二阶段我们提出了主题类别定义(TCD)模块，该模块在语料$D_{i1}$的基础上，细致分析不同主题类别在写作板块特征及其内容侧重点的差异性，通过这种方法，我们能够对每个主题类别进行精准的特征描述，并将其归纳整理成一个结构化的主题类别特征库，以便在给定成年人文本$x$进行主题类别分类后，准确地匹配到与之相对应的主题特征。在第三阶段，我们构建儿童新闻文本生成的提示，以指导LLM依据成年人新闻文本的核心内容，生成适合儿童阅读的新闻文本。最后我们对新生成的文本$D'_{i2}$与儿童新闻文本语料$D_{i1}$从多维度进行对比分析，以验证我们方法的有效性。TNC-LLM的整体框架图如图5所示。

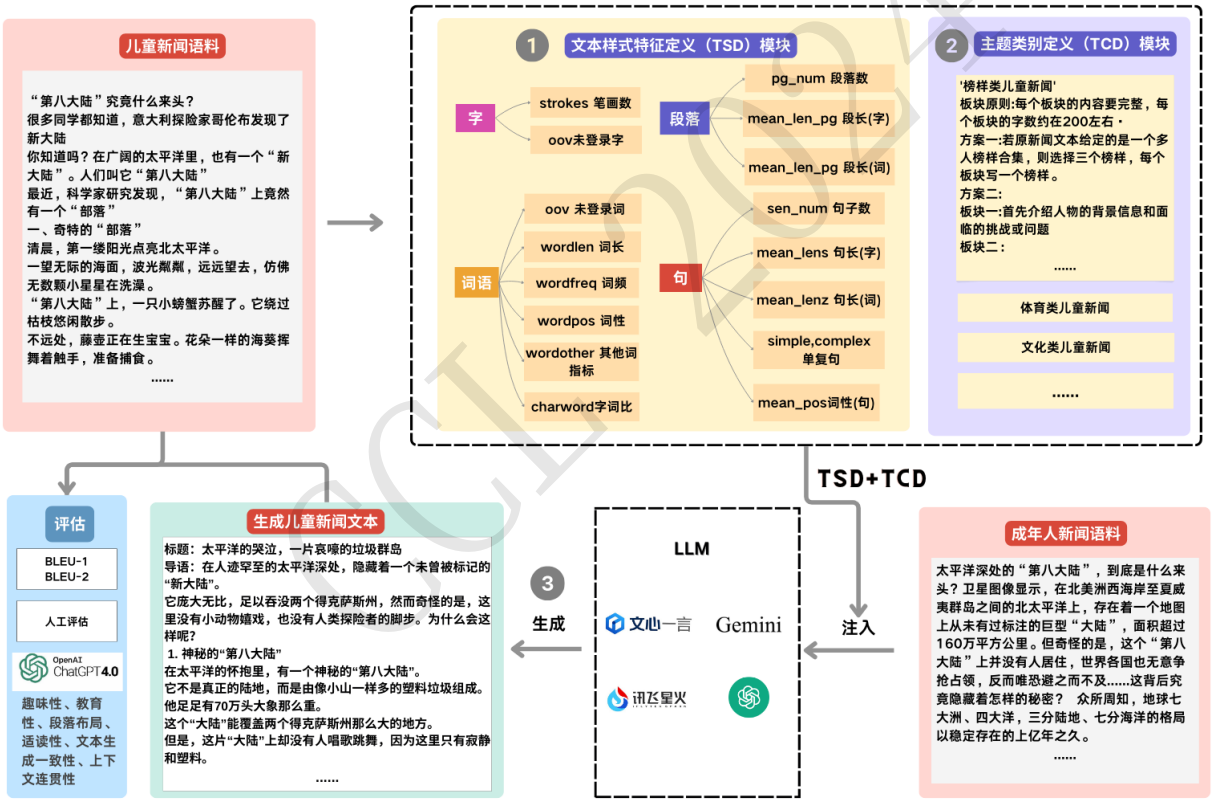



图 5. 基于主题的儿童新闻篇章结构框架（TNC-LLM）图

## 4.2 文本样式定义(TSD)模块

如图6我们设计了一个专为儿童新闻文本定制的文本样式定义（TSD）模块。该模块从字、词、句及段落四个级别统计了$D_{i1}$的文本样式，并给出了具体的样式定义。TSD模块能够为大型语言模型提供准确的文本样式提示，有利于在成年人新闻内容转化处理上，降低文本难度从而适配儿童的阅读理解能力，还可以优化文章版式布局。

**字** 在研究汉字结构的复杂性时，笔画数是一个重要的量化指标，通常笔画越多，记忆和认

知的难度越大。本研究统计分析了儿童新闻语料库中汉字的笔画分布，以了解其复杂度特征。同时，考虑到儿童对常见汉字的熟悉度，我们分析了汉字的熟悉度和未登录字的比例，这两个维度对理解儿童汉字认知具有重要意义。

笔画数：综合了(Sung et al., 2015)和(孙汉银, 1992)分类方法，本研究将笔画数小于或等于8的汉字划分为低笔画字，将笔画数大于16的汉字划分为高笔画字，其余的汉字则被归类为中等笔画字(吴思远et al., 2020)。汉字熟悉度：为了准确评估未登录字的比例，本研究参考了六个不同的字表进行量化分析，这些字表包括《现代汉语语料库字频表》、《常用字表》以及该表中前500、前1000、前1500和前2000位的高频字集合。

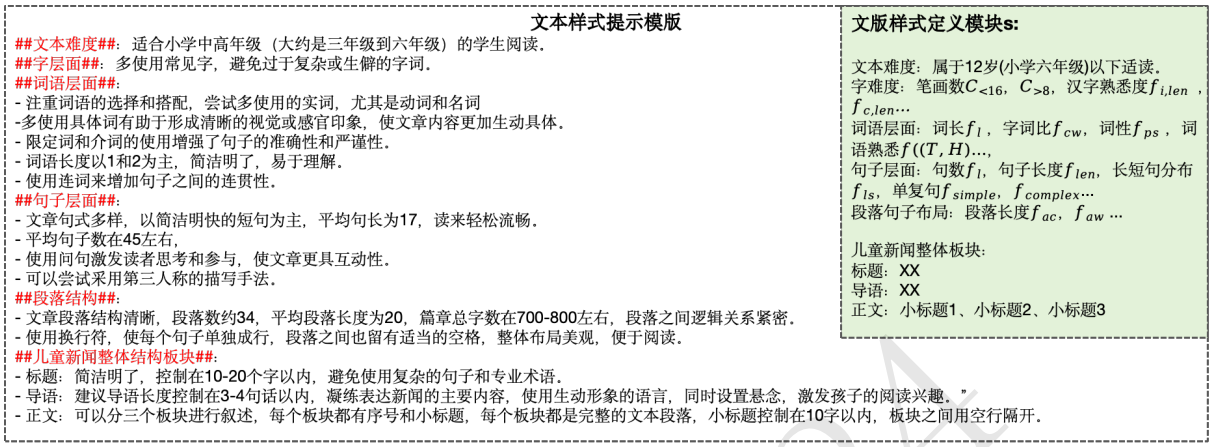



图 6. 文本样式定义模块示意图

**词语** 词汇是构建文本的基础，对于塑造语言的特色具有极其重要的作用，(Sung et al., 2015)从字词复杂度、词语长度、词语频率、词语计数和词汇多样性等多个维度评估词汇复杂度。通过对文本的词汇进行定量分析可以帮助我们更好地理解文本的特征和含义，本文中从词长、字词比、词频、词性等多个维度来分析。

词长：在本研究中，我们专注于从汉字的角度来量化词长，即通过计算词语中所含汉字的数量来评估其长度。为了量化词长，我们采用公式1，其中$f_l$表示平均词长；$argTop$是对加权长度进行求和以得到平均值；$U$ 表示集合的并集操作，在这里是将所有不同长度单词的加权长度合并起来；$K_l$是权重因子，它可能与长度$l$ 的单词在文本中的相对重要性或频率有关，$I_l$ 是长度为$l$的单词的计数，$nw$ 表示文本中不同单词长度的最大值(Tao et al., 2024)。

$$f_l = argTop\left(\bigcup_{i\in\{1,2,\ldots,\mathrm{nw}\}} K_l \cdot I_l\left(len\left(w_i\right)\right)\right) \tag{1}$$

字词比：字词比是指文本中汉字总数与词语总数之间的比率，这一概念由(张卫国, 2006)提出。通过这个量化指标，我们可以更准确地评估和比较不同文本的词汇复杂度。词频：我们应用了jieba2中的Cut函数库，对文本$D_{i1}$进行标记，以获得每个单词的词性注释。词类主要分为实词、虚词和否定词等。通过对文章中的词汇进行分类和统计分析，可以推断出词性的分布。词语熟悉度：首先通过jieba对$D_{i1}$中文本进行分词，分别统计词出现在《现代汉语语料库词频表》的前500、前1000、前1500 与前2000 高频词的数量。

**句** 在中文文本中，句子是由一系列词构成的表达单元，它表达了完整的思想或信息，句子的长度和难度等信息可以帮助我们了解文本的结构和组织方式，可以评估文本的阅读难度等级和适用人群，本文从句子数以及句子的长度、长短句、单复句等来分析文本的特征和结构特点。

句子数：我们以句号、问号或感叹号等标点符号结束来统计文本中的句子数。长短句：我们评估每个句子的长度，并根据20个字的阈值将其分类为短句或长句，其中$I_{ls}()$是长句和短句计数的函数。如果文章主要包含长句，则被归类为长句较多；相反，它被归类为有更多的短

句。如果长短句数量大致相等，则认为文章是长句和短句的组合(Tao et al., 2024)。

$$f_{ls} = i \in \{1, 2, \ldots, n_s\} argmax \{I_{ls}(s_i)\} \tag{2}$$

单复句：复句通常包含更多的信息和复杂的句子结构，统计复句的数量可以帮助我们评估文本的复杂性。较高的复句比例可能意味着文本内容较为复杂，需要读者有较高的理解能力。为了计算文本中的单复句，用句法分析器构建句子的句法树，这些树表示句子的结构和成分。通过分析这些树，我们可以识别主句和从句，从而判断一个句子是单句还是复句。

**段落** 段落长度对文本的可读性和理解性有重要影响。较短的段落通常更容易阅读和理解，而较长的段落可能需要读者更多的注意力和努力(Peter, 1996)。分析段落长度可以帮助改进文本的布局和格式，使其更适应目标读者。通过计算文本段落数和平均段落字数和平均段落词数，可以了解文本的结构和组织方式。设$T$为给定的文本，它由一系列段落$P = (p_1, p_2, ..., p_m)$组成，其中$m$是段落的总数。每个段落$p_i$包含$n_i$ 个字。段落的平均字数计算公式如下，其中，$|p_i|_{chars}$表示段落$p_i$中的字符数（包括空格和标点符号）。统计结果在附录B.3。

$$f_{ac} = \frac{\sum_{i=1}^{m} |p_i|_{chars}}{m} \tag{3}$$

### 4.3 主题类别定义(TCD)模块

成年人新闻遵循标准的报道模式，包括标题、导语、正文、背景和结尾。这些报道使用正式语言和专业术语，以及复杂句法，确保信息权威准确。它们常用被动语态以显客观中立，并深入分析事件原因、影响和后果，涵盖政治、经济、社会和文化等关键领域，旨在激发读者批判性思考，提供多元视角和深度讨论。报道采用“倒金字塔”结构，优先展示核心信息，并通过逻辑和过渡词保持文章连贯性，同时融入背景资料和专家意见。

相比之下，儿童新闻的结构更为简化，主要由标题、导语和正文三个部分组成。内容上侧重于儿童日常生活相关主题，如学校、社会和自然等。目的在于简化信息，使其易于理解，重点在于传授基础知识和道德教育，而不是深入分析。儿童新闻在文本表现上追求形象、声音和色彩的丰富性，强调故事性和生动性，运用各种修辞手法来提升语言的吸引力和文本的可读性(Alon-Tirosh, 2014)。在段落布局上，儿童新闻采用分节式结构，每篇新闻都分为三个板块，每个板块讨论一个特定主题或情景，通过具体情境的描述吸引儿童的注意力，传递积极正面的主题和价值观，助力儿童树立正确的人生观和价值观，增强社会责任感。

在深入分析儿童新闻的撰写特点后，我们发现不同主题类别的儿童新闻在内容呈现上各有侧重。因此在本文中我们将儿童新闻划分为不同的主题类别，以便更精确地满足儿童受众的需求和兴趣。该模块针对特定的主题，采用适合该主题的写作风格和内容策略，确保新闻报道既有趣又具有教育意义，同时能够吸引儿童的注意力并促进他们的理解。通过这种方式，儿童新闻能够更有效地传递信息，激发儿童的好奇心和学习兴趣。如图7是主题类别定义模块中环境保护类儿童新闻的提示模版，附录C.1中为完整的主题类别定义（TCD）模块。

环境保护类儿童新闻：

板块原则：每个板块的内容要完整，每个板块的字数约在200左右，每个板块有小标题
板块一：首先介绍背景和问题的存在，让读者了解目前面临的挑战或问题的严重性。
板块二：深入探讨问题带来的具体影响或潜在危害。这一部分旨在增强读者对问题紧迫性的认识。
板块三：提出解决问题的方法、已经采取的措施或呼吁采取的行动。这部分旨在激励读者或社会采取行动，对问题进行缓解或解决。

环境类儿童新闻的撰写特点：
互动性和参与感：通过提问或呼吁读者思考，激发儿童对环境问题的兴趣和参与感，如通过展示积极的改变，激发他们对保护环境的兴趣和责任感。
科普性：在故事化的叙述中融入科学知识和环保理念，既娱乐又教育，帮助儿童读者在轻松愉快的阅读过程中学习到有关环境保护的知识。
生动的比喻和描绘：通过将复杂的环境问题比作人类熟悉的情景(如“地球之肺”生病)，使儿童更容易理解和感受到问题的紧迫性。
简化的数据和事实：使用容易理解的数据(如“每过8秒，就有足球场大小的森林消失”)来说明问题，避免了复杂和抽象的统计数据，使信息更加亲近儿童的认知水平。

图 7. 主题类别定义模块中环境保护类儿童新闻的提示模版


(c) 2024 中国中文信息学会计算语言学专业委员会

# 5 实验与结果

## 5.1 实验设计

为了评估我们框架的全面性和适用性，我们选取了四种当前领先的大型语言模型（LLM）作为实验模型。这些模型包括：GPT-3.5-turbo-16k、讯飞星火-v3.5、ERNIE-3.5、Gemini-1.0-pro。通过它们，我们可以全面地评估我们的框架在不同场景下的表现和适用性。我们研究了在Zero-Shot和One-Shot学习场景下，利用文本样式定义模块和主题类别定义模块的组合提示，四个不同的模型生成儿童新闻内容的能力。其中所有模型的温度均为0.5。在Zero-Shot场景中，模型没有接受到任何具体的示例或指导，需要直接根据给定的模块和成年人版新闻文本内容生成儿童新闻文本。Zero-Shot结合文本样式定义模块s和主题类别定义模块c的提示模版在附录A.3。而在One-Shot场景中，模型首先提供一个成年人版新闻文本和对应的儿童新闻文本示例，然后基于这个示例结合给定模块组合生成新的儿童新闻内容。通过这种设置，我们可以评估模型在不同学习条件下的文本生成性能，以及它们理解和应用给定模块提示的能力。此外，我们将采用不提供任何模块提示的情况作为实验的基准线。One-Shot结合文本样式定义模块s和主题类别定义模块c的提示模版在附录A.2中。

## 5.2 实验数据与评估指标

我们使用的数据集包括两个部分：$D_{i1}$中包含534篇长篇儿童新闻文本，以及$D_{i1}$中与长篇儿童新闻文本主题内容相对应的534篇成年人新闻文本。我们的评估主要分为两个方面，一方面是在验证内容和主题的保留，另一方面是评估文本的趣味性、教育性、段落布局、适读性、文本生成一致性、上下文连贯性。其中在内容相似性上的评估使用BLEU-n(n=1,2)，在趣味性、教育性、段落布局、适读性、文本生成一致性、上下文连贯性方面使用GPT-4来进行评估(Chang et al., 2024)。为确保GPT-4评估结果的准确性和可信度，我们从每个评估模型中随机抽取了50篇文本，总计200篇，邀请四位语言学硕士来进行背靠背人工评审。这一步骤旨在检验GPT-4的评分是否与人工评估人员的专业判断保持一致。通过这种方法，我们能够对GPT-4的评估效果进行严格的验证，并据此调整和优化我们的评估流程。完整的GPT-4评估提示在附录A.1中，考虑LLM生成文本的随机性，我们将每个类型的文本传输3次，并计算最终的平均结果，其中实验结果统计为百分制。

## 5.3 实验结果及分析

在多个自然语言处理任务中，例如事件抽取、对话生成和文本摘要，大型语言模型在只有少量样本（Few-shot）的情况下表现出色。在本研究中，将TNC模块分别与Zero-Shot和One-Shot学习方法相结合，进行了相关实验。如图表1是在BLEU-n下的评估结果，其中TNC表示使用文本样式定义（TSD）模块和主题类别定义（TCD）模块的组合。在Zero-Shot实验中，Gemini-1.0-pro模型在BLEU-1值在40.9%，显示出生成文本和原文本内容上中等程度的一致性。与GPT-3.5-turbo-16k的得分相近，在加入我们的基于主题的儿童新闻篇章结构框架（TNC-LLM）结果显示，这两个模型的分数有显著提升，尤其在GPT-3.5-turbo-16k的BLEU-1上达到了50.35%，这表明TNC框架对于提升模型性能是非常有效的。讯飞星火-v3.5模型即使没有TNC框架其BLEU-1也达到了44.38%，可能因其在生成儿童文本方面有优势。这四个模型在加入TNC后BLEU值都有所提升。在One-Shot实验中，One-Shot情况下的分数普遍高于Zero-Shot情况，在多个模型中应用了TNC框架的版本普遍比未应用的版本表现更好，其中讯飞星火-v3.5在加入TNC后，BLEU-n值达到最高值，再次表明TNC框架是一种有效的性能提升手段。也同时反映出不同的模型在处理儿童新闻文本方面有不同的内在能力。

为了验证GPT-4评估结果的准确性，我们采取了随机抽样的方法进行人工复审。在对200篇随机选取的评估结果进行人工审核后，我们发现大约有168篇的评估与人工判断相符，其余32篇的人工评审结果与GPT-4评审结果差值统计详情在附录B.2，这一结果表明GPT-4的评估具有较高的准确度。基于这一发现，我们决定将GPT-4的评估结果作为我们分析的重要参考依据。如图表2是使用GPT-4在六个维度上的评估结果，结果统计为等比百分制。如图9为各个模型的实验结果在趣味性等六个维度的雷达图。在Zero-Shot条件下，结果显示，TNC框架的引入在多数维度上提升了模型的表现。这一现象在结合GPT-3.5-turbo-16k模型时尤为明显，其中TNC+GPT-3.5-turbo-16k在趣味性、教育性和段落布局上的得分相较于未加入TNC框架的原始模型提升较为明显，分别提高了4.22、10.66和9.39。此外，TNC+讯飞星火-v3.5在趣味


(c) 2024 中国中文信息学会计算语言学专业委员会

性、教育性、段落布局、适读性上表现最为出色，再次表明该模型在生成儿童新闻文本方面具有优势。在One-Shot条件下，大多数模型在实验设置下都取得了更高的分数，尤其是在趣味性、教育性和段落布局方面。TNC+GPT-3.5-turbo-16k在多个维度上取得了最高分，其中趣味性得到显著的提升，TNC+讯飞星火-v3.5在教育性和段落布局上均取得最高分，这一结果表明TNC框架不仅能够提升模型的文本生成质量，还能够在儿童教育中发挥出更大的潜力。TNC框架通过引入特定的教育性内容，优化了文本结构，使得生成的文本更加符合教育目的，并提高了儿童的阅读兴趣和理解能力。在表2所展示的数据中，对于一致性和连贯性这两个指标，不同模型的表现并没有显著差异。然而，当我们将TNC模块注入系统后，我们发现这两个指标有轻微的提升。这表明TNC模块对于提高文本的一致性和连贯性具有积极作用。然而，在ERNIE-3.5模型中，我们的方法在一致性和连贯性上的表现出现了轻微的下降。经过深入分析，我们推测这一现象的原因可能是由于将两个模块组合的提示注入到ERNIE-3.5模型中，这可能对ERNIE-3.5模型的生成过程施加了不必要的限制和约束，从而影响了其性能。

| 模型名称 | Zero-Shot | | One-Shot | |
|---|---|---|---|---|
| | **BLEU-1** | **BLEU-2** | **BLEU-1** | **BLEU-2** |
| Gemini-1.0-pro | 40.9 | 33.74 | 42.15 | 35.15 |
| **TNC+Gemini-1.0-pro** | 43.03 | 35.8 | 45.95 | 37.63 |
| GPT-3.5-turbo-16k | 41.19 | 33.89 | 43.98 | 35.92 |
| **TNC+GPT-3.5-turbo-16k** | **50.35** | **41.11** | 51.48 | 41.71 |
| 讯飞星火-v3.5 | 44.38 | 36.5 | 44.89 | 37.05 |
| **TNC+讯飞星火-v3.5** | 46.06 | 38.04 | **52.07** | **45.41** |
| ERNIE-3.5 | 41.02 | 33.33 | 42.53 | 34.27 |
| **TNC+ERNIE-3.5** | 44.73 | 35.38 | 48.41 | 40.32 |

表 1. 生成文本在BLEU-n指标下的评估结果(%)

| | 模型名称 | 趣味性 | 教育性 | 段落布局 | 适读性 | 一致性 | 连贯性 |
|---|---|---|---|---|---|---|---|
| Zero-Shot | Gemini-1.0-pro | 67.84 | 75.36 | 72.06 | 80.14 | 81.48 | 85.64 |
| | **TNC+Gemini-1.0-pro** | 70.06 | 79.34 | 79.16 | 81.32 | 81.68 | 85.58 |
| | GPT-3.5-turbo-16k | 65.82 | 70.32 | 70.88 | 79.52 | 81.26 | 85.49 |
| | **TNC+GPT-3.5-turbo-16k** | 69.84 | 80.98 | 80.27 | 81.52 | 81.49 | **85.88** |
| | 讯飞星火-v3.5 | 72.76 | 79.3 | 78.35 | 81.39 | 81.35 | 85.27 |
| | **TNC+讯飞星火-v3.5** | **75.77** | **81.31** | **81.72** | **81.56** | 81.2 | 85.46 |
| | ERNIE-3.5 | 63.74 | 72.01 | 69.83 | 79.26 | **81.66** | 85.43 |
| | **TNC+ERNIE-3.5** | 67.34 | 73.82 | 77.22 | 81.49 | 81.31 | 85.35 |
| One-Shot | Gemini-1.0-pro | 68.14 | 79.43 | 78.85 | 81.3 | 81.31 | 85.52 |
| | **TNC+Gemini-1.0-pro** | 71.74 | 80.45 | 81.94 | 82.09 | 81.34 | 85.39 |
| | GPT-3.5-turbo-16k | 67.33 | 75.05 | 77.6 | 80.08 | 82.05 | 85.47 |
| | **TNC+GPT-3.5-turbo-16k** | **75.56** | 82.91 | 82.04 | **82.54** | **82.79** | **86.36** |
| | 讯飞星火-v3.5 | 75.04 | 80.22 | 81.33 | 81.17 | 81.65 | 85.64 |
| | **TNC+讯飞星火-v3.5** | 75.53 | **83.28** | **82.78** | 82.27 | 82.77 | 85.56 |
| | ERNIE-3.5 | 65.67 | 73.21 | 78.23 | 81.13 | 81.94 | 85.91 |
| | **TNC+ERNIE-3.5** | 68.78 | 75.29 | 79.46 | 81.87 | 81.09 | 85.35 |

表 2. GPT-4在六个维度上对生成文本的评估结果

综合来看，TNC框架的引入在提高大型语言模型在儿童新闻文本生成任务上的表现方面起到了关键作用。无论是在零样本还是一样本条件下，引入TNC框架的模型普遍优于未加入该框架的模型。这可能是因为TNC框架通过提供更细粒度的控制和优化，使得模型能够更好地理解和适应儿童新闻文本生成任务的需求，从而在生成文本的内容一致、趣味性、教育性、段落布局、适读性和连贯性等多个维度上实现了显著提升。

### 5.4 消融研究

为了探究不同模块对儿童新闻生成的影响，我们进行了消融实验，分别评估了文本样式定义(TSD)模块和主题类别定义(TCD)模块的独立作用对儿童新闻文本生成的影响。消融实验的评估结果如表3和4。根据实验结果以及附录B.1中图9，可以看出在Zero-Shot情况下，我们观察到单独使用TSD模块与单独使用TCD模块在结合Gemini-1.0-pro、GPT-3.5-turbo-16k和ERNIE-3.5模型时，在多数评估指标表现相似。但对于讯飞星火-v3.5模型，单独


(c) 2024 中国中文信息学会计算语言学专业委员会

使用TSD模块在BLEU评分上取得了较高的分数，表明其在词汇重叠度和流畅性方面具有优势。在趣味性、教育性和段落布局这些维度上，单独使用TCD模块的表现更佳，这可能是因为TCD模块在内容生成方面更加注重创意和教育价值的融入。在One-Shot情况下，独立使用TCD模块时，ERNIE-3.5和讯飞星火-v3.5模型在BLEU值、趣味性、教育性上的表现显著优于单独使用TSD模块。这说明TCD模块借助于示例，能更有效地增强文本的适用性和教育价值。

| | Zero-Shot | | One-Shot | |
|---|---|---|---|---|
| **模型名称** | **BLEU-1** | **BLEU-2** | **BLEU-1** | **BLEU-2** |
| TCD+Gemini-1.0-pro | 41.59 | 32.41 | 42.78 | 35.47 |
| TSD+Gemini-1.0-pro | 42.73 | 35.72 | 42.64 | 35.42 |
| **TNC+Gemini-1.0-pro** | 43.03 | 35.8 | 45.95 | 37.63 |
| TCD+GPT-3.5-turbo-16k | 43.73 | 35.32 | 44.17 | 35.97 |
| TSD+GPT-3.5-turbo-16k | 43 | 35.37 | 43.44 | 35.56 |
| **TNC+GPT-3.5-turbo-16k** | **50.35** | **41.11** | 51.48 | 41.71 |
| TCD+ERNIE-3.5 | 42.97 | 34.96 | 45.59 | 36.37 |
| TSD+ERNIE-3.5 | 42.07 | 34.57 | 42.45 | 34.67 |
| **TNC+ERNIE-3.5** | 44.73 | 35.38 | 48.41 | 40.32 |
| TCD+讯飞星火-v3.5 | 43.19 | 35.4 | 46.29 | 38.57 |
| TSD+讯飞星火-v3.5 | 45.47 | 37.38 | 45.17 | 37.1 |
| **TNC+讯飞星火-v3.5** | 46.06 | 38.04 | **52.07** | **45.41** |

表 3. 消融实验生成文本在BLEU-n指标下的评估结果

| | **模型名称** | **趣味性** | **教育性** | **段落布局** | **适读性** | **一致性** | **连贯性** |
|---|---|---|---|---|---|---|---|
| Zero-Shot | TCD+Gemini-1.0-pro | 67.79 | 75.36 | 74.69 | 81.47 | 81.48 | 85.83 |
| | TSD+Gemini-1.0-pro | 68.62 | 80.52 | 78.32 | 81.27 | 81.35 | 85.69 |
| | **TNC+Gemini-1.0-pro** | 70.06 | 79.34 | 79.16 | 81.32 | 81.68 | 85.58 |
| | TCD+GPT-3.5-turbo-16k | 68.62 | 80.52 | 78.32 | 81.27 | 81.35 | 85.69 |
| | TSD+GPT-3.5-turbo-16k | 67.79 | 75.36 | 74.69 | 81.47 | 81.48 | 85.83 |
| | **TNC+GPT-3.5-turbo-16k** | 69.84 | 80.98 | 80.27 | 81.52 | 81.49 | **85.88** |
| | TCD+ERNIE-3.5 | 65.28 | 73.3 | 75.1 | 81.34 | **81.84** | 85.83 |
| | TSD+ERNIE-3.5 | 64.42 | 73.27 | 73.47 | 81.26 | 81.61 | 85.85 |
| | **TNC+ERNIE-3.5** | 67.34 | 73.82 | 77.22 | 81.49 | 81.31 | 85.35 |
| | TCD+讯飞星火-v3.5 | 75.35 | **81.37** | 80.59 | 81.37 | 81.73 | 85.52 |
| | TSD+讯飞星火-v3.5 | 72.54 | 80.33 | 78.77 | 81.52 | 81.37 | 85.67 |
| | **TNC+讯飞星火-v3.5** | **75.77** | 81.31 | **81.72** | **81.56** | 81.2 | 85.46 |
| One-Shot | TCD+Gemini-1.0-pro | 69.65 | 79.97 | 79.52 | 81.45 | 81.08 | 85.39 |
| | TSD+Gemini-1.0-pro | 68.69 | 79.38 | 78.35 | 80.98 | 81.01 | 85.16 |
| | **TNC+Gemini-1.0-pro** | 71.74 | 80.45 | 81.94 | 82.09 | 81.34 | 85.39 |
| | TCD+GPT-3.5-turbo-16k | 69.6 | 81.55 | 79.59 | 81.51 | 81.5 | 85.43 |
| | TSD+GPT-3.5-turbo-16k | 67.73 | 75.24 | 76.83 | 81.43 | 81.26 | 85.48 |
| | **TNC+GPT-3.5-turbo-16k** | **75.56** | 82.91 | 82.04 | **82.54** | **82.79** | **86.36** |
| | TCD+ERNIE-3.5 | 67.83 | 74.36 | 78.91 | 81.25 | 81.41 | 85.67 |
| | TSD+ERNIE-3.5 | 65.63 | 73.34 | 78.31 | 81.37 | 81.5 | 85.38 |
| | **TNC+ERNIE-3.5** | 68.78 | 75.29 | 79.46 | 81.87 | 81.09 | 85.35 |
| | TCD+讯飞星火-v3.5 | 75.41 | 82.28 | 82.01 | 81.33 | 81.54 | 85.63 |
| | TSD+讯飞星火-v3.5 | 74.86 | 81.18 | 81.8 | 81.11 | 81.29 | 85.04 |
| | **TNC+讯飞星火-v3.5** | 75.53 | **83.28** | **82.78** | 82.27 | 82.77 | 85.56 |

表 4. 消融实验在GPT-4在六个维度上对生成文本的评估结果

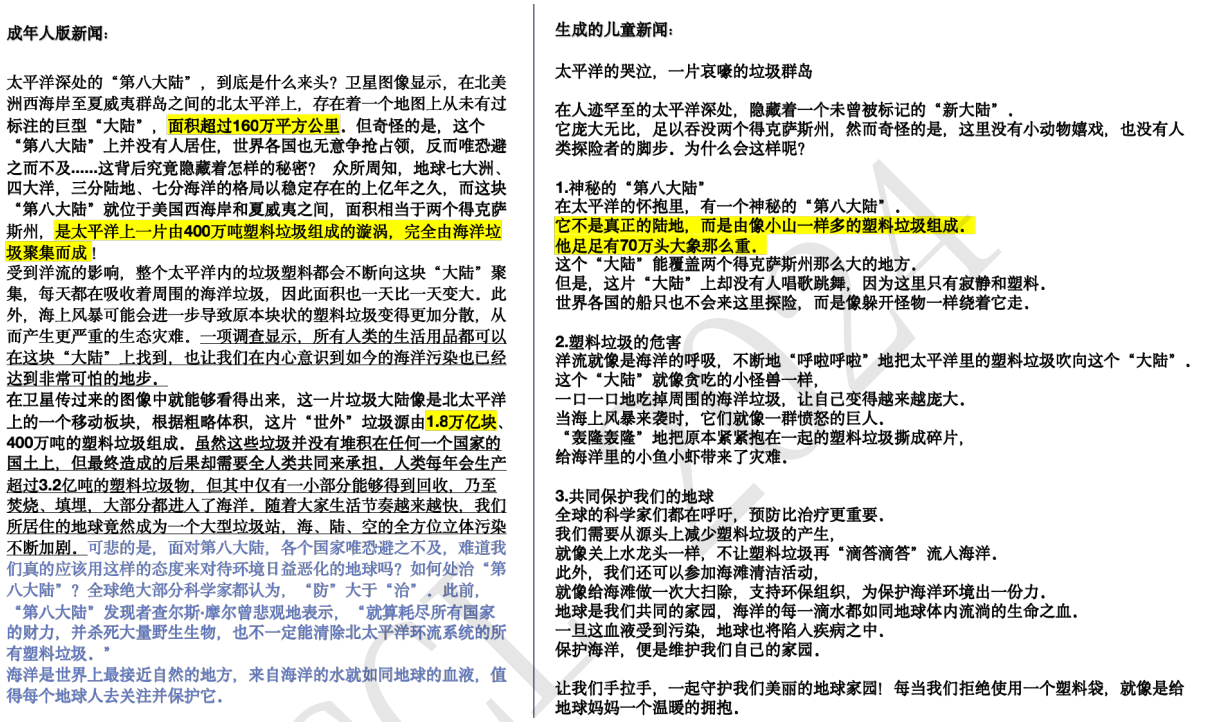
TCD和TSD模块各有所长，它们结合不同模型在实验条件下展现出各自的优势。TCD模块在内容创意和教育价值方面表现突出，而TSD模块在提升文本的流畅性和BLEU评分方面更具优势。综合实验的结果，将TCD模块与TSD模块结合使用，能够显著提升儿童新闻生成框架的整体性能。这种组合不仅在文本的流畅性和一致性方面表现出色，同时也在趣味性、教育性、段落布局和适读性等关键维度上达到了更高的标准。通过这种协同作用，我们能够生成更符合儿童阅读习惯和认知水平的新闻内容，从而更好地满足青少年儿童的新闻阅读需求。因此，TCD与TSD的结合不仅是一种有效的策略，更是实现高质量儿童新闻生成的关键。

### 5.5 案例分析

如图8右侧文本是将我们的TSD模块和TCD模块注入到讯飞星火v-3.5根据左侧的成年人新


(c) 2024 中国中文信息学会计算语言学专业委员会

闻内容生成的儿童版新闻文本。观察生成的儿童新闻，我们可以发现其排版布局整洁有序，段落短小精悍，且广泛使用简短的句子。相比之下，成年人版新闻在语言风格上更为严肃和直接，频繁使用如“垃圾塑料”、“生态灾难”和“可怕的生态污染”等词汇，以此突出问题的紧迫性。与此相对照，儿童新闻倾向于运用生动形象的比喻，例如“像小山一样多的塑料垃圾”和“像贪吃的小怪兽”，为了更好地抓住孩子们的注意力，文本中巧妙地融入了互动性强的语言元素，比如使用拟声词“滴答滴答”和“轰隆轰隆”等，以吸引儿童的注意力。成年人版新闻在文中高亮部分提供了详尽的数据和事实依据，如“面积超过160万平方公里”、“400万吨塑料垃圾”和“1.8万亿块塑料垃圾”，而儿童新闻则通过类比——如“足足有70万头大象那么重”，来帮助儿童把握问题的规模。在文章的左侧画线文字部分，成年人版深入探讨了问题的严重性，这部分内容在制作儿童版新闻时被刻意省略。在文章结尾（浅色部分），成年人版客观地阐述了问题的严重性，目的在于提升公众对环境问题的认识，同时隐含地对政策制定者和行业实践提出批评。相较之下，儿童版新闻更注重培养环保意识，通过一系列生动的比喻和形象化的描述，激发孩子们对保护地球的责任感和行动力。



**成年人版新闻：**

太平洋深处的“第八大陆”，到底是什么来头？卫星图像显示，在北美洲西海岸至夏威夷群岛之间的北太平洋上，存在着一个地图上从未有过标注的巨型“大陆”，面积超过160万平方公里。但奇怪的是，这个“第八大陆”上并没有人居住，世界各国也无意争抢占领，反而唯恐避之而不及……这背后究竟隐藏着怎样的秘密？ 众所周知，地球七大洲、四大洋，三分陆地、七分海洋的格局以稳定存在的上亿年之久，而这块“第八大陆”就位于美国西海岸和夏威夷之间，面积相当于两个得克萨斯州，是太平洋上一片由400万吨塑料垃圾组成的漩涡，完全由海洋垃圾聚集而成！
受到洋流的影响，整个太平洋内的垃圾塑料都会不断向这块“大陆”聚集，每天都在吸收着周围的海洋垃圾，因此面积也一天比一天变大。此外，海上风暴可能会进一步导致原本块状的塑料垃圾变得更加分散，从而产生更严重的生态灾难。一项调查显示，所有人类的生活用品都可以在这块“大陆”上找到，也让我们在内心意识到如今的海洋污染也已经达到非常可怕的地步。
在卫星传过来的图像中就能够看得出来，这一片垃圾大陆像是北太平洋上的一个移动板块，根据粗略体积，这片“世外”垃圾源由1.8万亿块、400万吨的塑料垃圾组成。虽然这些垃圾并没有堆积在任何一个国家的国土上，但最终造成的后果却需要全人类共同来承担。人类每年会生产超过3.2亿吨的塑料垃圾物，但其中仅有一小部分能够得到回收，乃至焚烧、填埋，大部分都进入了海洋。随着大家生活节奏越来越快，我们所居住的地球竟然成为一个大型垃圾站，海、陆、空的全方位立体污染不断加剧。可悲的是，面对第八大陆，各个国家唯恐避之不及，难道我们真的应该用这样的态度来对待环境日益恶化的地球吗？如何处治“第八大陆”？全球绝大部分科学家都认为，“防”大于“治”。此前，“第八大陆”发现者查尔斯·摩尔曾悲观地表示，“就算耗尽所有国家的财力，并杀死大量野生生物，也不一定能清除北太平洋环流系统的所有塑料垃圾。”
海洋是世界上最接近自然的地方，来自海洋的水就如同地球的血液，值得每个地球人去关注并保护它。

**生成的儿童新闻：**

太平洋的哭泣，一片哀嚎的垃圾群岛

在人迹罕至的太平洋深处，隐藏着一个未曾被标记的“新大陆”。
它庞大无比，足以吞没两个得克萨斯州，然而奇怪的是，这里没有小动物嬉戏，也没有人类探险者的脚步。为什么会这样呢？

1.神秘的“第八大陆”
在太平洋的怀抱里，有一个神秘的“第八大陆”。
它不是真正的陆地，而是由像小山一样多的塑料垃圾组成。
他足足有70万头大象那么重。
这个“大陆”能覆盖两个得克萨斯州那么大的地方。
但是，这片“大陆”上却没有人唱歌跳舞，因为这里只有寂静和塑料。
世界各国的船只也不会来这里探险，而是像躲开怪物一样绕着它走。

2.塑料垃圾的危害
洋流就像是海洋的呼吸，不断地“呼啦呼啦”地把太平洋里的塑料垃圾吹向这个“大陆”。
这个“大陆”就像贪吃的小怪兽一样，
一口一口地吃掉周围的海洋垃圾，让自己变得越来越庞大。
当海上风暴来袭时，它们就像一群愤怒的巨人，
“轰隆轰隆”地把原本紧紧抱在一起的塑料垃圾撕成碎片，
给海洋里的小鱼小虾带来了灾难。

3.共同保护我们的地球
全球的科学家们都在呼吁，预防比治疗更重要。
我们需要从源头上减少塑料垃圾的产生，
就像关上水龙头一样，不让塑料垃圾再“滴答滴答”流入海洋。
此外，我们还可以参加海滩清洁活动，
就像给海滩做一次大扫除，支持环保组织，为保护海洋环境出一份力。
地球是我们共同的家园，海洋的每一滴水都如同地球体内流淌的生命之血。
一旦这血液受到污染，地球也将陷入疾病之中。
保护海洋，便是维护我们自己的家园。

让我们手拉手，一起守护我们美丽的地球家园！每当我们拒绝使用一个塑料袋，就像是给地球妈妈一个温暖的拥抱。

图 8. 成年人版新闻与讯飞星火v-3.5模型生成的儿童新闻示例

## 6 结论与未来工作

在本研究中，我们提出了一种创新性的基于主题的框架儿童新闻篇章结构框架（TNC-LLM），这是大型语言模型在儿童新闻领域的首次系统性应用。本框架的核心由两个关键模块构成：文本样式定义（TSD）模块和主题类别定义（TCD）模块。经过实验我们发现TCD模块在内容创意和教育价值方面表现突出，而TSD模块在提升文本的流畅性和BLEU评分方面更具优势。综合实验的结果，TNC框架通过提供更细粒度的控制和优化，使得模型能够更好地理解和适应儿童新闻生成任务，从而在生成文本的趣味性等多个维度上实现了显著提升。同时我们发现讯飞星火-v3.5在趣味性、教育性表现最为出色，表明该模型在生成面向儿童教育文本方面具有优势。尽管我们的框架已经取得了相对成功，但我们认识到，未来的工作仍有进一步完善和提升的空间。我们计划在未来的研究中更加关注知识补充和整合方面的工作。由于儿童新闻往往需要综合多篇成年人新闻的信息，如何有效地提炼和融合这些知识以生成既准确又富有教育意义的儿童新闻，将成为我们未来研究的重点。我们期望通过这些努力，能够进一步提升儿童新闻的教育价值，使其成为儿童学习和成长过程中的有益资源。


(c) 2024 中国中文信息学会计算语言学专业委员会

## 参考文献

Alon-Tirosh. 2014. If i was making the news": What do children want from news. *Participations*, 11(1):108–129.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3), mar.

Banghao Chen, Zhaofeng Zhang, Nicolas Langrené, and Shengxin Zhu. 2023. Unleashing the potential of prompt engineering in large language models: a comprehensive review.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and play language models: A simple approach to controlled text generation.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation.

Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. 2023. Challenges and applications of large language models.

Michael Lebowitz. 1985. Story-telling as planning and learning. *Poetics*, 14:483–502.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: A simple approach to sentiment and style transfer.

Danyang Liu, Juntao Li, Meng-Hsuan Yu, Ziming Huang, Gongshen Liu, Dongyan Zhao, and Rui Yan. 2020. A character-centric neural model for automated story generation. In *AAAI Conference on Artificial Intelligence*.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity.

N. Newman. 2024. Journalism, media, and technology trends and predictions 2024. Retrieved from https://policycommons.net/artifacts/11300794/journalism-media-and-technology-trends-and-predictions-2024/12186166/. Accessed on 17 Apr 2024, CID: 20.500.12592/fbg7g5r.

Kyeongman Park, Nakyeong Yang, and Kyomin Jung. 2023. Longstory: Coherent, complete and length controlled long story generation.

Pelletier. 2004. Action, consciousness and theory of mind: Children's ability to coordinate story characters' actions and thoughts. *Early Education Development*, 15(1):5–22.

Peter. 1996. Working memory, intelligence and reading ability in children. *Personality and Individual Differences*, 21(6):1007–1020.

Jing Qian, Li Dong, Yelong Shen, Furu Wei, and Weizhu Chen. 2022. Controllable natural language generation with contrastive prefixes.

Yao Ting Sung, Ju Ling Chen, Ji Her Cha, Hou Chiang Tseng, Tao Hsing Chang, and Kuo En Chang. 2015. Constructing and validating readability models: the method of integrating multilevel linguistic features with machine learning. *Behavior Research Methods*.

Zhen Tao, Dinghao Xi, Zhiyu Li, Liumin Tang, and Wei Xu. 2024. Cat-llm: Prompting large language models with text style definition for chinese article-style transfer. *ArXiv*, abs/2401.05707.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Jingjing Xu, Xu Sun, Qi Zeng, Xiaodong Zhang, Xuancheng Ren, Houfeng Wang, and Wenjie Li. 2018. Unpaired sentiment-to-sentiment translation: A cycled reinforcement learning approach. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 979–988, Melbourne, Australia, July. Association for Computational Linguistics.


(c) 2024 中国中文信息学会计算语言学专业委员会

Peng Xu, Mostofa Patwary, Mohammad Shoeybi, Raul Puri, Pascale Fung, Anima Anandkumar, and Bryan Catanzaro. 2020. MEGATRON-CNTRL: Controllable story generation with external knowledge using large-scale language models. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2831–2845, Online, November. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models.

Yang Zhong, Chao Jiang, Wei Xu, and Junyi Jessy Li. 2020. Discourse level factors for sentence deletion in text simplification.

吴思远, 于东, and 江新. 2020. 汉语文本可读性特征体系构建和效度验证. 世界汉语教学, 34(1):81–97, 1.

孙汉银. 1992. 中文易懂性公式. 硕士学位论文, 北京师范大学.

张卫国. 2006. 阅读:覆盖率,识读率和字词比. 语言文字应用, (3):8.

皮亚杰. 1981. 儿童心理学. 商务印书馆, 北京.

饶芳芳. 2023. 具身认知视角下儿童科普读物设计研究——以《小小牛顿幼儿馆steam》系列为例. 上海包装, (11):137–139.

CCL 2024

# A

## A.1 GPT-4评估生成文本的提示模版

你是个儿童新闻文本编辑专家，请客观的的将模型生成文本的质量参照儿童新闻语料，从以下四个维度：趣味性、教育性、段落布局、适读性来进行参考评估，文本风格越接近原文本分数越高，从1（最差）到10（最好）进行评分。其中每个维度的具体信息如下：

**趣味性**

故事元素：文本是否包含吸引人的故事，如有趣的角色、激动人心的冒险或富有教育意义的情节。

修辞手法：文本是否使用了比喻、拟人等修辞手法来增强描述的生动性和形象性。

想象力和创造力：文本是否鼓励儿童进行想象，提供创造性思考的机会，或者激发他们的创造性表达。

**教育性**

知识传授：文本是否介绍了新的知识点，如科学发现、历史事件或文化习俗。

生活技能：文本是否提供了实用的信息或建议，帮助儿童学习生活中的技能，如健康饮食、安全意识等。

道德价值：文本是否讨论了道德价值观念，如诚实、勇敢、同情心等，并以儿童易于理解的方式呈现。

**段落布局**

结构清晰：是否一篇文章分三个小标题板块进行叙述，每个段落是否有明确的开始和结束，以及是否围绕单一主题或想法组织内容。

视觉呈现：段落是否通过标题、子标题等视觉元素来增强可读性和吸引力。

**适读性**

语言难度：文本使用的词汇和句子结构是否适合12岁（小学六年级）以下儿童的年龄和阅读水平。

信息量：文本是否提供了适量的信息，避免过于简单或过于复杂，确保儿童能够理解和吸收。

**文本生成一致性**

主题一致性：文本是否在整篇文章中保持对主题的一致关注，避免偏离主题。

风格一致性：文本的语气、风格和语言特点是否全文保持一致，没有突兀的变化。

**上下文连贯性**

内部逻辑：文本中的事件、描述和论点是否逻辑上连贯，形成有意义的整体。

连接词使用：是否恰当使用连接词和过渡短语来连接句子和段落，确保文本的流畅性。

上下文关联：文本中的信息是否与上下文相关联，确保读者能够理解每个部分在整个故事或讨论中的位置和作用。

**原文本**={儿童新闻语料}

**需要评估的文本**={模型生成的文本}

## A.2 One-Shot下的文版样式定义模块s和主题类别定义模块提示模版

示例example:

成年人版新闻文本：XXX
儿童版新闻文本：XXX

成年人版新闻文本$D_{i2}$： “央视网消息：北京时间今天（6月15日）13时30分，我国在太原卫星发射中心使用长征二号丁运载火箭，成功将吉林一号高分06A星等41颗卫星发射升空，卫星顺利进入预定轨道，发射任务获得圆满成功。该批卫星主要用于提供商业遥感服务及相关技术验证。此次任务是长征系列运载火箭的第476次飞行。41位“乘客” “一箭多星”再创纪录。“1箭41星”再次刷新了中国航天“一箭多星”纪录，41颗卫星就像是41位“乘客”，发射前它们要被合理地安排在火箭的整流罩里；发射后，它们是分组“下车”、依次分离，这就给发射任务带来了不小的挑战。太原卫星发射中心的科技人员从转运、吊装、卫星安装、星罩合体等各个


(c) 2024 中国中文信息学会计算语言学专业委员会

环节，精心准备、精心组织、精心实施，确保了发射任务圆满成功.....”

提示： 你是一位资深的儿童文学专家，现在有一个根据成年人版新闻文本$D_{i2}$内容来生成一篇儿童版的新闻，请参考以下两个特征模块：文本样式定义模块s和主题类别定义模块c。请参考以上示例example。（注：确保内容真实准确，避免误导儿童，选择儿童感兴趣的话题和事件。）

文版样式定义模块s:
文本难度：属于12岁（小学六年级）以下适读。
字难度：笔画数$C_{<16}$，$C_{>8}$，汉字熟悉度$f_{i,len}$，$f_{c,len}$
词语层面：词长$f_l$，字词比$f_{cw}$，词性$f_{ps}$，词语熟悉度$f(T,H)$
段落句子布局：句数$f_{sl}$，句子长度$f_{len}$,$f_{lw}$，长短句分布$f_{ls}$，单复句$f_{simple}$，$f_{complex}$，段落长度$f_{ac}$，$f_{aw}$
儿童新闻整体板块：标题：XX、导语：XX、正文：小标题1、小标题2、小标题3。

主题类别定义模块c:
该文本可以写成XX类别的新闻，该类别的儿童新闻主题特征如下:$c_i$

### A.3 Zero-Shot下的文版样式定义模块s和主题类别定义模块提示模版

成年人版新闻文本$D_{i2}$: “央视网消息：北京时间今天（6月15日）13时30分，我国在太原卫星发射中心使用长征二号丁运载火箭，成功将吉林一号高分06A星等41颗卫星发射升空，卫星顺利进入预定轨道，发射任务获得圆满成功。该批卫星主要用于提供商业遥感服务及相关技术验证。此次任务是长征系列运载火箭的第476次飞行。41位“乘客” “一箭多星”再创纪录。“1箭41星”再次刷新了中国航天“一箭多星”纪录，41颗卫星就像是41位“乘客”，发射前它们要被合理地安排在火箭的整流罩里；发射后，它们是分组“下车”、依次分离，这就给发射任务带来了不小的挑战。太原卫星发射中心的科技人员从转运、吊装、卫星安装、星罩合体等各个环节，精心准备、精心组织、精心实施，确保了发射任务圆满成功.....”

提示： 你是一位资深的儿童文学专家，现在有一个根据成年人版新闻文本$D_{i2}$内容来生成一篇儿童版的新闻，请参考以下两个特征模块：文本样式定义模块s和主题类别定义模块c。（注：确保内容真实准确，避免误导儿童，选择儿童感兴趣的话题和事件。）

文版样式定义模块s:
文本难度：属于12岁（小学六年级）以下适读。
字难度：笔画数$C_{<16}$，$C_{>8}$，汉字熟悉度$f_{i,len}$，$f_{c,len}$
词语层面：词长$f_l$，字词比$f_{cw}$，词性$f_{ps}$，词语熟悉度$f(T,H)$
段落句子布局：句数$f_{sl}$，句子长度$f_{len}$,$f_{lw}$，长短句分布$f_{ls}$，单复句$f_{simple}$，$f_{complex}$，段落长度$f_{ac}$，$f_{aw}$
儿童新闻整体板块：标题：XX、导语：XX、正文：小标题1、小标题2、小标题3。

主题类别定义模块c:
该文本可以写成XX类别的新闻，该类别的儿童新闻主题特征如下:$c_i$

# B

## B.1 多维度实验评估结果雷达图

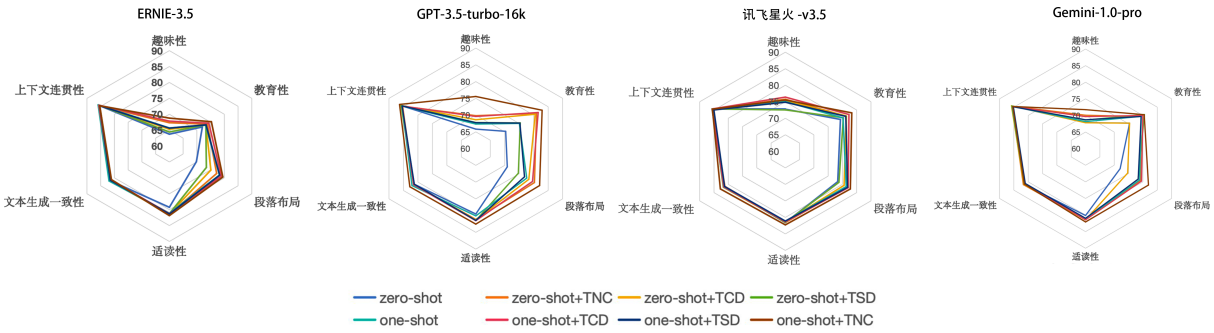



图 9. 多维度实验评估结果雷达图

## B.2 32篇人工评估与GPT-4评估不符统计详情

| 段落布局 | 趣味性 | 教育性 | 可读性 |
|---|---|---|---|
| -2 | +1 | +1 | +1 |
| -1 | +1 | +1 | +1 |
| -1 | +1 | +2 | -1 |
| -2 | -1 | +1 | -1 |
| -2 | +1 | -1 | -1 |
| +1 | -1 | -1 | +1 |
| -2 | +2 | - | -2 |
| -2 | +1 | - | - |
| +2 | +2 | - | - |
| - | +2 | - | - |

表 5. 人工评估与GPT-4评估对比（+2代表人工评估比GPT-4评估在该维度多2分）

CCL 2024

## B.3 TSD实验结果

| | | | |
|---|---|---|---|
| 平均笔画数（argv_bh） | 7.42 | 字词比（charword_token） | 1.57 |
| 最大笔画数(maxi_bh) | 19.38 | 词长为1的字数(len1) | 212.07 |
| 最小笔画数(mini_bh) | 1.00 | 词长为2的字数(len2) | 169.11 |
| 笔画数小于8的字数(low8_bh) | 423.03 | 词长为3的字数(len3) | 20.32 |
| 笔画数大于16的字数(hi16_bh) | 9.50 | 词长为4的字数(len4) | 6.75 |
| 平均字数(var_docfreq_word) | 787.65 | 词长大于4的字数(len4_) | 1.05 |
| 不在字频表高频500的字数(oov_corf500_char) | 200.64 | 否定词数(neg) | 14.01 |
| 不在字频表高频1000的字数(oov_corf1000_char) | 99.18 | 成语数(idiom) | 2.18 |
| 不在字频表高频1500的字数(oov_corf1500_char) | 53.30 | 实词数(cont) | 318.72 |
| 不在字频表高频2000的字数(oov_corf2000_char) | 30.03 | 虚词数(func) | 95.79 |
| 不在常见字表的字数(oov_commonchar) | 6.58 | 形容词（a） | 23.68 |
| 汉字难度等级均值(argv_dl_tokenchar_newhsk) | 2.53 | 其他名词修饰语（b） | 1.25 |
| 出现在常见字表1000的字数(common_10) | 405.26 | 连词（c） | 5.46 |
| 出现在常见字表2000的字数(common_20) | 574.78 | 副词（d） | 37.16 |
| 出现在常见字表3500的字数(common_35) | 637.32 | 感叹词（e） | 0.87 |
| 不在词频表高频500的词数(oov_corf500_word) | 207.18 | 成语（i） | 4.33 |
| 不在词频表高频1000的词数(oov_corf1000_word) | 175.71 | 数字（m） | 20.07 |
| 不在词频表高频1500的词数(oov_corf1500_word) | 154.33 | 普通名词（n） | 93.13 |
| 不在词频表高频2000的词数(oov_corf2000_word) | 140.83 | 指示名词（nd） | 8.29 |
| 疑问代词数（PronWhat） | 2.17 | 人名（nh） | 1.42 |
| 并列连词（binlieConj） | 3.13 | 地理名词（ns） | 9.28 |
| 递进连词（dijinConj） | 0.64 | 时间名词（nt） | 5.96 |
| 假设连词（jissheConj） | 2.61 | 其他专有名词（nz） | 0.29 |
| 让步连词（rangbuConj） | 1.51 | 拟声词（o） | 0.93 |
| 顺承连词（shunchengConj） | 0.16 | 介词（p） | 12.43 |
| 条件连词（tiaojianConj） | 1.01 | 数量词（q） | 13.26 |
| 因果连词（yinguoConj） | 1.46 | 代词（r） | 24.28 |
| 转折连词（zhuanzheConj） | 1.64 | 助动词（u） | 38.93 |
| 连词总数（ConjAll） | 12.17 | 动词（v） | 110.32 |
| 动词短语数（VP_num） | 62.39 | 标点符号（wp） | 117.78 |
| 名词短语数（NP_num） | 103.01 | 描写性词（z） | 1.72 |
| 副词短语数（ADVP_num） | 49.64 | 代词数（PronAll） | 10.39 |
| 介词短语数（PP_num） | 10.37 | 指示代词数（PronDemo） | 4.51 |
| 形容词短语数（ADJP_num） | 9.68 | 人称代词数（PronPers） | 3.71 |
| 句子数（sent） | 45.24 | 第一人称代词数（PronPers_Firs） | 0.88 |
| 单句数（easy_sent） | 13.07 | 第二人称代词数（PronPers_Secd） | 0.63 |
| 复句数（complex_sent） | 32.17 | 第三人称代词数（PronPers_Thir） | 2.21 |
| 平均句子长度(字)（mean_lenz） | 14.38 | 段落数（pg_num） | 34.99 |
| 最大句子长度(字)（max_lenz） | 35.89 | 平均段落长度(字)（mean_lenz_pg） | 18.58 |
| 最小句子长度(字)（min_lenz） | 2.46 | 平均段落长度(词)（mean_lenw_pg） | 15.12 |
| 平均句子长度(词)（mean_lenw） | 11.94 | | |
| 最大句子长度(词)（max_lenw） | 29.17 | | |
| 最小句子长度(词)（min_lenw） | 2.54 | | |

表 6. TSD模块的统计结果

CCL 2024

# C

## C.1 主题类别定义(TCD)板块

榜样类儿童新闻:

板块原则:每个板块的内容要完整，每个板块的字数约在200左右，每个板块有小标题。

方案一:若原新闻文本给定的是一个多人榜样合集，则选择三个榜样，每个板块写一个榜样。

方案二:

板块一:首先介绍人物的背景信息和面临的挑战或问题

板块二:接着描述人物是如何通过创新或特殊的方法解决问题的。

板块三:最后展示人物的努力和创新带来的积极成果和影响。

榜样类儿童新闻的撰写特点:

注重人物故事:通过讲述个人的故事和经历，使新闻更加生动和吸引人。

强调正面价值:通过展示主人公面对困难的勇气、持之以恒的精神传递正面的价值观和榜样的力量。

启示与反思:通过榜样的故事，引发读者的启示和反思

寓教于乐:在吸引儿童注意力的同时，传递积极的价值观和行为模范，鼓励儿童学习和模仿。

可读性强:通过生动的故事情节和具体的人物形象，要善于使用一些增强语言生动效果的修辞手法，比如比喻，使得新闻内容富有趣味性和吸引力，容易被儿童接受和理解。

环境保护类儿童新闻:

板块原则

每个板块的内容要完整，每个板块的字数约在200左右，每个板块有小标题

板块一：首先介绍背景和问题的存在，让读者了解目前面临的挑战或问题的严重性。

板块二：接着，深入探讨问题带来的具体影响或潜在危害。这一部分旨在增强读者对问题紧迫性的认识。

板块三：最后，提出解决问题的方法、已经采取的措施或呼吁采取的行动。这部分旨在激励读者或社会采取行动，对问题进行缓解或解决。

环境类儿童新闻的撰写特点:

互动性和参与感：通过提问或呼吁读者思考，激发儿童对环境问题的兴趣和参与感。如通过展示积极的改变激发他们对保护环境的兴趣和责任感。

科普性：在故事化的叙述中融入科学知识和环保理念，既娱乐又教育，帮助儿童读者在轻松愉快的阅读过程中学习到有关环境保护的知识。

生动的比喻和描绘：通过将复杂的环境问题比作人类熟悉的情景（如\地球之肺"生病），使儿童更容易理解和感受到问题的紧迫性。

简化的数据和事实：使用容易理解的数据（如\每过8秒，就有足球场大小的森林消失"）来说明问题，避免了复杂和抽象的统计数据，使信息更加亲近儿童的认知水平。

经济类儿童新闻:

板块原则:

每个板块的内容要完整，每个板块的字数约在200左右，每个板块有小标题

板块一：以生活中熟悉的事物或现象作为引入，或通过讲述一个具体的例子来吸引儿童的兴趣激发儿童的兴趣。

板块二：在吸引了读者的注意力之后，文本转而介绍了背景信息和当前的问题，这样做能帮助儿童理解实验的背景和目的。

板块三：探讨解决方案，引发了关于其好处或坏处的讨论。这一部分旨在激发儿童对结果的好奇心，并鼓励他们思考和讨论。


(c) 2024 中国中文信息学会计算语言学专业委员会

经济类儿童新闻的撰写特点：

语言简洁生动：使用易于理解和生动的语言来解释经济概念，避免了复杂的术语和概念，使儿童能够更容易地理解。

情境化教学：通过创造一个具体的情境或故事来介绍经济概念，使抽象的经济理论变得具体化，更容易吸引儿童的注意力。

鼓励思考和讨论：提出了问题和可能的后果，鼓励儿童进行思考和讨论，培养他们的批判性思维能力。

关联现实生活：通过与儿童的现实生活相关联的例子来解释经济概念，或介绍具体的行动和政策，强调理论知识在实际生活中的应用，使得内容更加贴近儿童的生活经验，增加了学习的实用价值。

科学类儿童新闻:

板块原则

每个板块的内容要完整，每个板块的字数约在200左右，每个板块有小标题

板块一：以引人入胜的方式提出一些日常生活中的问题或现象，激发儿童的好奇心。

板块二：介绍针对提出的问题的科学实验过程和结果 / 科技或方法/ 科技产品的功能和作用

板块三：给出了实验的结论，介绍这些发现或发明的实际应用，这不仅满足了儿童对问题答案的好奇心，还向他们展示了科学研究对日常生活的影响。

科技类儿童新闻的撰写特点：

科学知识与价值观教育相结合：不仅传递科学知识，还通过探索的故事传达了勇于探索未知、追求知识的价值观

趣味性： 文章选题有趣，语言风格幽默，使得复杂的科学问题变得易于理解和亲近，适合儿童的阅读习惯。

教育性： 通过介绍真实的科学研究和实验，文章传达了科学探索的方法和精神，鼓励儿童保持好奇心，培养他们的科学思维

实用性： 文章通过介绍科学研究的实际应用，如解决日常生活中的小问题，让儿童了解科学知识的实用价值，增强他们学习科学的动力。

社会类儿童新闻:

板块原则

每个板块的内容要完整，每个板块的字数约在200左右，每个板块有小标题

板块一：通过生动的描述引入事件的背景和情境，

板块二：介绍事件中的关键角色和行动，详细叙述事件的过程和影响。

板块三：揭示故事背后的深层意义，强调正面价值观和教育意义。

社会类儿童新闻的撰写特点：

生动形象：通过具体的场景描写和比喻，让儿童能够直观感受到故事情境，增强阅读的吸引力。

情感共鸣：通过故事化的方式，让儿童能够与新闻中的人物或事件产生情感上的共鸣，从而激发他们的同情心和正义感。

启发式教育：在报道中穿插启发性问题或思考，引导儿童思考事件的原因、结果以及可能的解决方案，培养他们的批判性思维和问题解决能力。

行动倡议：提供简单、实际的行动建议或参与方式，让儿童明白他们也能为社会问题做出贡献，增强他们的社会参与感和责任感。

文化多样性：展现不同文化和社会背景下的事件和人物，增进儿童对多元文化的认识和尊重，培养他们的全球视野和包容心态。

安全意识：在报道涉及风险或危机的事件时，强调安全措施和预防方法，教育儿童如何在类似情况下保护自己和他人。


(c) 2024 中国中文信息学会计算语言学专业委员会

体育类儿童新闻:

板块原则

每个板块的内容要完整，每个板块的字数约在200左右，每个板块有小标题

板块一：以一个独特、有趣的场景或事件作为开头，介绍事件的背景、地点、时间和一些吸引人的信息，为读者建立一个场景。

板块二：深入讲述事件的具体内容，包括人物、地点、特殊情境等，使读者能够更加详细地了解事件。

板块三：在故事的结尾，传达一个正面的信息或价值观，让读者从故事中获得启发。

体育类儿童新闻的撰写特点:

道德和价值观的强调：通过故事传递积极的价值观，如勤奋、公平竞争、团队精神、坚持不懈、社区参与、多元文化的重要性等，鼓励儿童积极参与体育活动，培养团队精神和坚持不懈的精神

情感丰富有趣：通过描述人物、情境和细节，激发读者的情感共鸣，增加故事的吸引力。

互动性：鼓励读者参与和思考，如通过提问或建议读者想象自己在场景中的行为。

易于理解的规则解释：对于体育项目的规则和术语，用简单易懂的语言进行解释，帮助儿童更好地理解比赛，增加他们的参与感。

健康和安全意识的普及：在报道中强调运动的安全意识和健康习惯，教育儿童如何在参与体育活动时保护自己，避免受伤。

文化类儿童新闻:

板块原则

每个板块的内容要完整，每个板块的字数约在200左右，每个板块有小标题

板块一：通过一个引人入胜的故事介绍文化背景或历史起源。这部分可以通过一个具体、生动的事例或者角色引入主题。

板块二：详细阐述文章的主题内容，通过具体的例子或故事来展开讨论。

板块三：最后一部分对文章进行总结，并提供更广泛的视角或思考，有时也可以提出问题，引导读者进行进一步的思考。

文化类儿童新闻的撰写特点:

故事化的叙述：使用简洁、生动的语言描述复杂的概念，通过讲故事的方式介绍文化现象或历史背景，增加文章的吸引力和记忆点。

历史文化知识的融入：通过故事介绍相关的历史文化知识，让儿童在阅读故事的同时增长知识。

鼓励探索和思考：通过展示文化现象的变迁和多样性，鼓励儿童读者思考文化的意义和城市的变化，激发他们对探索和学习的热情。、互动性：通过提问或小活动，让儿童参与到文化学习中来，例如，可以邀请他们想象自己是历史事件的见证者或参与者，或者设想自己如何将学到的文化知识应用到日常生活中。

政治类儿童新闻:

板块原则

每个板块的内容要完整，每个板块的字数约在200左右，每个板块有小标题

板块一：首先介绍事件的背景，为孩子们提供足够的信息来理解事件的重要性。例如，介绍基辛格为何要进行秘密访问中国，以及这次访问的历史背景。

板块二：详细叙述事件的经过，突出重要的转折点和细节。这部分内容帮助孩子们理解事件的发展过程，如基辛格的秘密行动细节、与中国官员的会谈等。

板块三：最后，阐述事件的结果和对未来的影响。这有助于孩子们理解事件的意义，例如中美关系的改善，以及基辛格对两国友谊所做出的贡献。

政治类儿童新闻撰写的特点:

简化复杂性：将复杂的政治事件简化，使用易于理解的语言和概念，同时保留事件的核心要素。

强调教育价值：突出事件的教育意义，例如友谊、和平、合作等正面价值观。

避免偏见：在描述政治事件时保持中立，避免传达任何形式的偏见或立场，让儿童自己形成看法。

鼓励思考和讨论：在新闻的结尾提出问题或思考点，鼓励孩子们思考讨论，培养他们的批判性思维能力

CCL 2024

# 基于对比学习和排名一致性的古代汉语翻译质量评估模型

**李怀明　邵艳秋　李炜***
北京语言大学，信息科学学院，
国家语言资源监测与研究平面媒体中心，
北京市海淀区学院路15号，100083
18811607325@163.com　yqshao163@163.com　liweitj47@blcu.edu.cn

## 摘要

当前，虽然机器翻译的自动评估技术已展现出良好的性能，但将它们应用于古代汉语到现代汉语的翻译场景时效果并不理想。一方面，这些传统方法能较好地比较质量差异较大的译文的好坏，但是在评估质量相差不大的译文时往往难以区分优劣。另一方面，古代汉语的省略和复杂句式常导致翻译过程中出现漏译现象，而传统评估指标往往会给这类较差的译文偏高的分数。在本文中，我们提出了一种基于对比学习和排名一致性的古代汉语到现代汉语的翻译质量评估模型（**CRATE**）。该模型通过确保语义相似度和匹配度的排名一致性捕捉译文质量的细粒度排名信息。另外，我们在使用对比学习方法训练译文跟原文的匹配模型时，将原文自身作为负样本，有效解决了传统评估指标在译文出现漏译情况下仍给出高评分的问题。为了证明我们模型的有效性，我们构建了高质量的古代汉语到现代汉语翻译的人工评分测试集。实验结果表明，我们的模型优于强大的基线，与人类评分取得了更显著的相关性。

**关键词：** 翻译评估；古代汉语；细粒度排名；对比学习

# Ancient Chinese translation quality evaluation model based on contrastive learning and ranking consistency

**Huaiming Li　Yanqiu Shao　Wei Li***
Information Science School, Beijing Language and Culture University,
Language Resources Monitoring and Research Center,
15 Xueyuan Road, HaiDian District, Beijing, 100083
18811607325@163.com　yqshao163@163.com　liweitj47@blcu.edu.cn

## Abstract

Currently, although automatic evaluation techniques for machine translation have shown good performance, their results are not ideal when applied to translation scenarios from ancient Chinese to modern Chinese. On the one hand, while these traditional methods can better compare translations with large differences in quality, they are often difficult to distinguish between translations with similar quality. On the other hand, omissions and complex sentence patterns in ancient Chinese often lead to missed translations during the translation process, and traditional evaluation metrics tend to give higher scores to such poor translations. In this paper, we propose a translation quality evaluation model from ancient Chinese to modern Chinese (**CRATE**) based on contrastive learning and ranking consistency. This model captures fine-grained ranking information of translation quality by ensuring ranking consistency for semantic similarity and matching. In addition, when using the contrastive learning method to train

* 通讯作者　Corresponding Author

第二十三届中国计算语言学大会论文集，第171页–第182页，太原，中国，2024年7月25日至28日。
卷1：主会论文
(c) 2024 中国中文信息学会计算语言学专业委员会

the matching model between the translation and the source text, the source text itself is used as a negative sample, which effectively solves the problem of traditional evaluation metrics still giving high scores when the translation is missing. To demonstrate the effectiveness of our model, we constructed high-quality human annotation test sets of ancient Chinese to modern Chinese translations. Experimental results show that our model outperforms strong baselines and achieves more significant correlations with human ratings.

**Keywords:** Translation evaluation , Ancient Chinese , Fine-grained ranking , Contrastive learning

## 1 引言

古代汉语作为一门历史深厚的语言体系，不仅承载了中国丰富的历史遗产和哲学思想，更是理解中华文明的核心基石。然而，随着时间的推移，语言的演变使古代汉语与现代汉语之间形成了显著的差异，从而对阅读和理解构成了重大的障碍。因此，将古代汉语精确地翻译成现代汉语显得尤为关键。鉴于这一挑战，开发一种科学且客观的古代汉语翻译质量评估方法变得尤为迫切。一个准确的评估方法不仅可以增强翻译的精确性和可靠性，而且能确保中国传统文化的精髓被准确传递。总之，古代汉语的翻译质量评估不仅对于语言学研究具有重要意义，同时对于文化传承和国际文化交流亦具有不可估量的价值。

早期的机器翻译主要依赖于人工评价。Church and Hovy (1993)从流利度、忠实度和理解力三个维度对翻译进行评估。Lommel et al. (2014)提出了多维度质量指标（MQM），详细列出了机器翻译容易出现的八种错误类型。这些评估方法至今仍对翻译评估产生影响。然而，人工评估不仅成本高，还易受评估人员主观性的影响。随着技术进步，机器翻译的自动评估方法应运而生。这些自动评估方法主要分为基于字符串匹配和基于语言模型两大类。基于字符串匹配的方法涵盖了编辑距离 (Su et al., 1992)、准确率 (Papineni et al., 2002)）和召回率 (Lin, 2004)等评估指标，因其操作简便，在机器翻译评估领域得到了广泛的应用。然而，这些指标主要依赖于字面上的词汇匹配，未能充分挖掘句子的深层语义，忽视了语言的多样性与复杂性。相反，基于语言模型的评估方法 (Zhang et al., 2020)能够通过理解句子的深层语义来有效处理同义的问题，在评价质量差异显著的译文时表现出色。但是，在区分质量相近的译文时，这类评估方法的效果往往不够理想。

此外，古代汉语是一种高度精练且紧凑的语言，常涉及省略等复杂语法结构，这使得将其翻译成现代汉语时容易出现漏译现象。在图1中，我们展示了几个翻译过程中出现漏译的示例。样例1中的“**妻子**”在古代汉语中表示“**妻子和孩子**”的含义，而现代汉语中通常仅指代“**妻子**”。样例2中的“**咸秩群祀**”和样例3中的“**考系廷尉**”过于简练，导致翻译模型难以准确捕捉其含义。这些因素均会导致漏译现象。然而，传统评估指标主要关注语义相似度，常对这类质量较低的译文给出偏高的分数。鉴于现有机器翻译评估指标在评价古代汉语翻译质量方面存在不足，本文旨在提出一种更精确的评估方法，以促进古代文本翻译研究的发展。

在本文中，我们提出了一个基于对比学习和排序一致性算法的古代汉语到现代汉语的翻译质量评估模型。该模型的基本思想是，当古代汉语有参考译文以及多个机翻译文时，这些机翻译文在与参考译文的语义相似度方面上的排名以及机翻译文在与古代汉语的翻译对匹配方面上的排名之间，应该表现出高度的一致性。我们通过确保机翻译文在语义相似度和匹配度两方面的**排名一致性**，来捕捉译文质量的**细粒度排名信息**。另外，我们在采用对比学习训练原文和译文的匹配度模型时，通过**使用原文自身作为负样本**，有效解决了传统评估方法对存在**漏译情况**的机器翻译给予过高评分的问题。此外，为了让对比学习在小批量数据中学习到更有效的信息，我们通过多种破坏并重构译文的方法构建了难负样本。

为了证明我们方法的有效性，我们构建了高质量的古代汉语翻译人工评估结果的测试集。实验结果表明，我们的模型性能超过了现有的主流翻译质量评估指标，与人类评分取得了更显著的相关性。此外，通过分析质量接近的译文与古代汉语之间距离的方差，证明了我们的方法

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

| | |
|---|---|
| 样例1 | **古代汉语：** 浑清素在公，妻子不免於饥寒。<br>**参考译文：** 郑浑为官的时候清正廉洁，家中妻子孩子经常忍饥挨饿。<br>**机翻译文：** 浑清素在公，妻子不免于饥寒。 |
| 样例2 | **古代汉语：** 六月庚子，初祀五岳四渎，咸秩群祀。<br>**参考译文：** 六月一日，前往祭祀五岳名山，百官都有秩序地参加祭礼。<br>**机翻译文：** 六月庚子日，最初的祭祀五岳四渎，咸秩群祀。 |
| 样例3 | **古代汉语：** 南阳圭泰尝以言迕指，考系廷尉。<br>**参考译文：** 南阳人圭泰曾在言语上违背了他们的指示，就被廷尉查考抓捕。<br>**机翻译文：** 南阳人圭泰曾用言语冒犯指，考系廷尉。 |

Figure 1: **古代汉语翻译过程中出现漏译情况的样例。**机翻译文来源于百度翻译[1]，漏译位置用不同颜色字体进行了标记。

能够提供更准确的细粒度排名信息。同时，我们单独对存在漏译情况的译文进行了实验分析，结果验证了我们在对比学习中将原文作为负样本的策略的有效性。

本文的主要贡献有：

- 我们提出了一个古代汉语翻译质量评估模型（CRATE），其通过确保机翻译文在语义相似度和匹配度两方面的排名一致性来捕捉细粒度的译文质量排名信息，改善了传统方法难以辨别质量相近译文优劣的问题。
- 在通过对比学习训练原文和译文的匹配模型时，我们将原文自身用作负样本，有效解决了传统评估指标在评价存在漏译现象的译文时倾向于给出过高分数的问题。
- 我们构建了高质量的古代汉语到现代汉语翻译的人工评估结果的数据集。实验结果表明，我们的模型的评估结果与人工注释结果高度相关，超过了当前的评估指标。

## 2 相关工作

机器翻译的自动评估方法最初基于字符串匹配，涵盖了编辑距离、准确率和召回率等方面。编辑距离的评估方法主要包括词误差率（WER）、位置无关错误率（PER）及翻译错误率（TER）。其中，WER (Su et al., 1992)通过计算机译文与参考译文之间的编辑距离来评估翻译质量，但对词序变化的惩罚过重，处理效果不佳。为了解决这一问题，PER (Tillmann et al., 1997)采用忽略词序的方法，而TER (Snover et al., 2006)则提出将词语移动的惩罚减至最低。基于准确率的BLEU (Papineni et al., 2002)和基于召回率的ROUGE (Lin, 2004)都是通过计算机译文与参考译文的n-gram重叠来进行评估。然而，这些指标主要关注词汇层面的重叠，未能充分体现句子的语义，对同义词或相似表达的处理尚显不足。为了弥补这一缺陷，METEOR (Banerjee and Lavie, 2005)引入了同义词词表，但其过分依赖于WordNet数据库，在中文翻译评估中的适用性有限。

随着预训练语言模型技术的快速发展，评估指标开始采用这些模型来理解句子的深层语义。例如，BERTscore (Zhang et al., 2020)利用语言模型来计算机翻译文跟参考译文之间的语义相似度，同时引入逆文本频率指数（IDF）来考虑词的重要程度。尽管这类指标在区分出质量相差较大的译文时表现出色，但在评估质量接近的译文方面效果不佳。当前，专门用于古代汉语到现代汉语翻译的质量评估方法相当有限。在这一领域，常用的评估方法仍然是基于字符串匹配的传统方法，如 (Li et al., 2018; Liu et al., 2020; Zhou and Si, 2023)在评估古代汉语翻译质量时仍然采用传统的BLEU评估指标。为了更好地满足古代汉语翻译的特殊需求，Yang et al. (2021)提出了针对古汉语翻译的自动评估指标DTE，其结合了BLEU和句子相似度评价方法的优点。然而，当机翻译文中出现漏译现象时，这些基于语义相似度的评估指标仍然倾向于给出偏高的评分。

[1] https://fanyi.baidu.com/


(c) 2024 中国中文信息学会计算语言学专业委员会

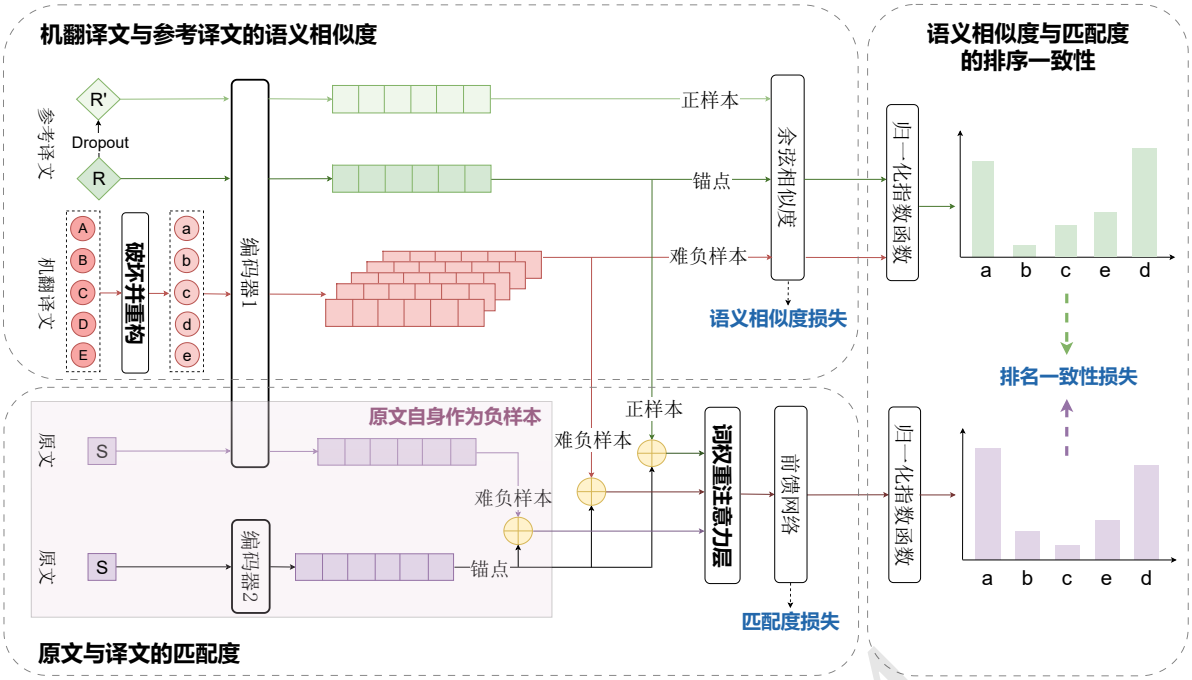



Figure 2: **古代汉语翻译评估模型CRATE的总体架构。** CRATE通过排名一致性来捕获细粒度的译文质量排名，并在对比学习中将原文作为负样本来解决漏译问题。该模型的总体训练目标是：(1)机翻译文和参考翻译之间语义相似度的对比学习损失；(2)机翻译文与原文的匹配方度的对比学习损失；(3)语义相似度与匹配度之间的排序一致性损失。

## 3 基于对比学习和排序一致性的古代汉语翻译质量评估模型

在本节中，我们介绍古代汉语翻译质量评估模型（CRATE）。CRATE的基本思想是通过确保语义相似度和匹配度两方面的排名的一致性来获得准确的细粒度的译文质量排名信息。具体而言，我们分别获取机翻译文在与参考译文的语义相似度方面上的排名以及机翻译文在与原文的匹配度方面上的排名，并确保它们之间的一致性。如图2所示，我们的方法由三部分组成：(1)机翻译文与参考译文的语义相似度（3.1节）；(2)原文与译文的匹配度（3.2节）；(3)语义相似度与匹配度的排名一致性（3.3节）。我们在3.4节介绍推理过程中译文质量分数的计算方法。

### 3.1 机翻译文与参考译文的语义相似度

该部分旨在计算机翻译文与参考译文之间的语义相似度分数。具体而言，我们通过对比学习的方法拉近语义相似的句子之间的距离，而让不同语义的句子距离更远，以此让编码器有效学习句子的向量表示。我们使用参考译文作为锚点。参考前人的工作 (Gao et al., 2021)，我们对参考译文应用dropout方法来生成语义相似的正样本。同时，我们采取调用语言模型API或者直接将古代汉语输入到翻译软件中，以获取多个机翻译文，之后再对它们进行破坏并重构的方法来构造难负样本。

**难负样本生成** 受到BART (Lewis et al., 2020)模型预训练过程中引入噪声的方式的启发，我们在构造负样本时通过不同的噪声随机破坏机翻译文，包括掩码、删除词语、随机词语替换、随机词语插入、置换词序等，再使用预训练语言模型BERT (Devlin et al., 2019)进行重构。该模型使用破坏并重构任务进行预训练，以便更好地重建被破坏的机翻译文。通过这种方法生成的是负样本更难，能够使模型在小批量数据中学习到更加有效的信息，并且可以更好地模拟翻译过程中可能遇到的各种错误。我们采用适当的比例来破坏文本，因为添加噪声太多会导致负样本没有包含有效的信息，添加噪声太少会导致负样本跟正样本过于接近，这样都不利于模型训练。我们在4.1节介绍了具体的细节。

之后，我们将这些处理好的锚点、正样本和负样本输入到预训练语言模型中，获得它们的句子向量表示，并通过计算余弦相似度来评估它们之间的语义相似性。在模型训练过程中，我


(c) 2024 中国中文信息学会计算语言学专业委员会

们以InfoNCE损失函数 (van den Oord et al., 2018)作为训练目标，因为其通过最大化正样本对的相对概率并最小化负样本对的相对概率，强化了模型对特征差异的敏感度。如式1所示：

$$\mathcal{L}_{similarity} = -log\frac{e^{sim(f_1(x_i),f_1(\overline{x}_i))/\tau_1}}{\sum_{j=1}^{N} e^{sim(f_1(x_i),f_1(x_j))/\tau_1}} \tag{1}$$

其中，$f_1(\cdot)$是通过预训练模型得到的句子向量表示，$x_i$ 是参考译文（锚点），$\overline{x}_i$ 是参考译文经过dropout后的文本（正样本），$x_j$ 是机翻译文经过破坏并重构后的文本（负样本），N是负样本的数量，$\tau_1$是温度超参数，$sim(f(x_i), f(x_j)) = \frac{f(x_i)^\top f(x_j)}{\|f(x_i)\| \cdot \|f(x_j)\|}$ 是余弦相似度。

### 3.2 原文与译文的匹配度

该部分旨在计算机翻译文与古代汉语之间的匹配分数。我们同样采用对比学习的方法进行训练，其中原文作为锚点，参考译文作为正样本，而破坏并重建的机翻译文作为负样本。另外，鉴于传统评估指标倾向于给那些存在漏译情况的译文较高的分数，我们还将原文自身作为了负样本，这样可以有效地降低这类质量较差的译文跟原文之间的匹配分数。

我们使用3.1节的预训练模型$f_1$来获取参考译文$x_i$的句子表示$h_i$（正样本）、经过破坏并重构之后的机翻译文$x_j$的句子表示$h_j$（负样本）以及原文$x_s$的句子表示$h_s$（负样本）。此外，我们使用guwenbert [2]作为预训练语言模型$f_2$来获得原文$x_s$的句子表示$h_{s'}$（锚点）。随后，我们分别将$h_i$、$h_j$、$h_s$与$h_{s'}$进行拼接。这个过程如式2所示：

$$h_{t+s'} = concat(h_t, h_{s'}),\ t\epsilon\{i, j, s\} \tag{2}$$

**词权重注意力层** Zhang et al. (2020)认为在评估译文的翻译质量时，词的重要程度并不相同，于是其使用IDF来计算不同词的权重信息。在本文中，我们应用注意力机制 (Vaswani et al., 2017)让模型集中处理更重要的语义信息。具体而言，我们使用权重向量$q$沿维度抽取拼接后的特征表示, 如式3所示：

$$\mathcal{V}_{t+s'} = q \cdot h_{t+s'},\ t\epsilon\{i, j, s\} \tag{3}$$

其中，$q$的值服从概率分布，且每行的概率的总和都为1。

随后，我们将抽取完的特征向量输入到前馈神经网络层，分别预测得到它们之间的匹配分数，如式4所示。

$$score_{t+s'} = \omega_1 ReLU(\omega_2 \mathcal{V}_{t+s'} + b_1) + b_2,\ t\epsilon\{i, j, s\} \tag{4}$$

其中，$\omega_1$、$\omega_2$、$b_1$ 和$b_2$ 是模型参数。

最后，我们通过对比学习最大化原文与参考译文之间的匹配分数，同时最小化原文与低质量译文之间的匹配分数。训练目标如图5所示：

$$\mathcal{L}_{match} = -log\frac{e^{score_{i+s'}/\tau_2}}{e^{score_{s+s'}/\tau_2} + \sum_{j=1}^{N} e^{score_{j+s'}/\tau_2}} \tag{5}$$

其中，$\tau_2$ 是温度超参数，N是负样本的数量。

### 3.3 语义相似度与匹配度的排名一致性

通过对比学习，我们能够有效地区分正样本跟负样本这种语义差距较大的样本。然而，模型并没有学习到负样本之间细粒度的差异，其仍然无法区分语义接近的样本。因此，我们通过确保语义相似度和翻译对匹配度两方面的排名的一致性来捕获译文的细粒度排名信息。具体而言，我们利用3.1节的语义相似度模型计算多个机翻译文跟参考译文的语义相似度的排名列表，利用3.2节的匹配模型计算多个机翻译文跟原文的匹配度的排名列表。

参考 Liu et al. (2023)的做法，我们利用ListNet (Cao et al., 2007)将排序列表看做概率分布。由于KL散度（Kullback-Leibler Divergence）不是对称的，所以我们通过最小化两个顶部概率分布（top one probability distributions）之间的JS散度（Jensen-Shannon divergence）来确保语义相似度和匹配度两方面的排名一致性，如式6所示。

[2] https://github.com/Ethan-yt/guwenbert


(c) 2024 中国中文信息学会计算语言学专业委员会

$$\begin{aligned}\mathcal{L}_{consistency} &= \sum_{i=1}^{N} JS(P_i \| Q_i) \\ &= \frac{1}{2}\sum_{i=1}^{N} KL(P_i \| \frac{P_i+Q_i}{2}) + \frac{1}{2}\sum_{i=1}^{N} KL(Q_i \| \frac{P_i+Q_i}{2}) \\ &= \frac{1}{2}\sum_{i=1}^{N} (P_i\ log(\frac{2P_i}{P_i+Q_i}) + Q_i\ log(\frac{2Q_i}{P_i+Q_i}))\end{aligned} \tag{6}$$

其中$P_i$和$Q_i$分别表示语义相似度列表和匹配分数列表的概率分布。

简而言之，$L_{similarity}$和$L_{match}$通过对比学习区分正负样本，忽略了负样本之间的差异性。$L_{consistency}$通过确保语义相似度以及匹配度两方面的排名一致性来捕获负样本之间细粒度的译文质量排名信息。我们模型的总的损失函数为：

$$\mathcal{L}_{local} = \mathcal{L}_{similarity} + \alpha\mathcal{L}_{match} + \beta\mathcal{L}_{consistency} \tag{7}$$

其中$\alpha$ 和$\beta$ 是超参数。

### 3.4 推理

在推理过程中，如果评估时无参考译文，我们需将机翻译文输入到编码器$f_1$中获取句子表示$h_j$，并将古代汉语输入到编码器$f_2$中获得句子表示$h_{s'}$。然后，将这两个句子表示进行拼接，并输入到注意力层和前馈神经网络中，计算出它们的匹配分数$score_{j+s'}$。

如果评估时有参考译文，我们则将参考译文和机翻译文输入到编码器$f_1$中获得$h_i$和$h_j$，并计算它们之间的余弦相似度。同样地，我们也需要计算机翻译文与古代汉语之间的匹配分数$score_{j+s'}$。最终的评估分数通过综合这两个分数来获得，公式如下：

$$score_{local} = \lambda score_{j+s'} + \mu sim(h_i, h_j) \tag{8}$$

其中$\lambda$ 和$\mu$ 是超参数。

## 4 实验部分

### 4.1 设置

参考 Ranasinghe et al. (2020)的工作，我们利用[CLS]标记对应的嵌入作为输入句子的表示。编码器在模型训练过程中更新参数。我们采用AdamW (Loshchilov and Hutter, 2019)作为优化器，并设置学习率为2e-5。此外，超参数$\alpha$和$\beta$的值设置为1，$\lambda$ 和$\mu$的值设置为0.5。我们设置负样本的数量为25，且调整温度超参数$\tau_1$ 和$\tau_2$为0.3（详情见4.3.4节）。另外，破坏机翻译文的比例我们设定为0.5。模型经过5个训练周期后，我们保留表现最佳的检查点，以便对测试集进行最终评估。

**数据** 本文的数据抽取自东北大学提供的一个包含古代汉语与现代汉语翻译对齐的数据集 [3]。我们随机抽取了20,550条数据对，其中20,000条数据对被划分为训练集。我们使用多个翻译模型得到古代汉语的机翻译文，包括gpt-3.5-turbo [4]，GLM-3-Turbo [5]，文心一言 [6]，wenyanwen-ancient-translate-to-modern [7]和百度翻译 [8]。对于模型，我们通过API调用方式获得译文；对于百度翻译，则直接输入古代汉语。余下的550条数据对经过筛选，剔除不合适的数据后，选取500条数据对中的古代汉语输入到wenyanwen-ancient-translate-to-modern来获取机翻译文。随后，我们根据原文和参考译文对这500条机翻译文进行人工评分，并将其作为测试集（UTS）。

[3] https://github.com/NiuTrans/Classical-Modern
[4] https://platform.openai.com/docs/models/gpt-3-5-turbo
[5] https://open.bigmodel.cn/usercenter/apikeys
[6] https://console.bce.baidu.com/qianfan/ais/console/applicationConsole/application
[7] https://huggingface.co/raynardj/wenyanwen-ancient-translate-to-modern
[8] https://fanyi.baidu.com/


(c) 2024 中国中文信息学会计算语言学专业委员会

**评估指标** 受到 Huang et al. (2023)工作的启发，我们通过计算模型预测与人工评分的皮尔逊相关系数(Pearson Correlation Coefficient)和斯皮尔曼等级相关系数(Spearman's Rank Correlation Coefficient)来评估模型性能。皮尔逊相关系数是关于两个随机变量之间的线性关系的统计度量，使用数据样本值本身进行计算，计算公式如式9所示。斯皮尔曼等级相关系数是随机变量之间单调关系的统计度量，使用数据样本排位位次值进行计算，计算公式如式10所示。

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{9}$$

其中，$\overline{x}$和$\overline{y}$是变量的均值。

$$\rho = \frac{\frac{1}{n}\sum_{i=1}^{n}((R(x_i) - \overline{R(x)}) \cdot (R(y_i) - \overline{R(y)}))}{\sqrt{(\frac{1}{n}\sum_{i=1}^{n}(R(x_i) - \overline{R(x)})^2) \cdot (\frac{1}{n}\sum_{i=1}^{n}(R(y_i) - \overline{R(y)})^2)}} \tag{10}$$

其中，$R(x_i)$和$R(y_i)$是变量$x$和$y$的位次，而$\overline{R(x)}$和$\overline{R(y)}$是对应排名的平均位次。

**基线** 我们将CRATE与其他评估指标进行了对比，包括BLEU、ROUGE和BERTscore。此外，我们还与古代汉语翻译质量评估指标DTE进行了比较。对于基于预训练模型的评估指标，我们尝试在$\text{BERT}_{\text{base}}$和$\text{RoBERTa}_{\text{base}}$ [9]上进行实验，我们的方法只更换$f_1$。另外，我们还在SimCSE[10]上进行了实验，通过直接计算机翻译文跟参考译文的语义相似度来评估译文质量。

### 4.2 实验结果

如表1所示，我们的方法在PCC和SRCC上与人类标注的相关性都是最高的，这验证了我们方法的有效性。在两个评估方法上，基于字符串的指标要明显差于基于预训练模型的指标，这证明前者确实只注意句子的表层词汇而忽视了深层语义分析。SimCSE的性能比BERTScore好很多，因为它通过对比学习获得了更准确的句子表示。此外，DTE的性能比SimCSE略好，因为其结合了BLEU以及句子相似度的优势，更贴合古文翻译的特点。值得注意的是，我们的方法无论基于BERT还是基于RoBERTa，在两个指标上都优于其他的基线。

| 预训练模型 | 方法 | PCC | SRCC |
|---|---|---|---|
| Non-BERT | BLEU | 0.4903 | 0.5087 |
| | ROUGE | 0.5125 | 0.5032 |
| $\text{BERT}_{\text{base}}$ | +BERTScore | 0.5971 | 0.6013 |
| | +SimCSE | 0.6530 | 0.6769 |
| | +DTE | 0.6927 | 0.6834 |
| | +CRATE (Ours) | **0.7153** | **0.7297** |
| $\text{RoBERTa}_{\text{base}}$ | +BERTScore | 0.6264 | 0.6427 |
| | +SimCSE | 0.6752 | 0.6618 |
| | +DTE | 0.6995 | 0.7019 |
| | +CRATE (Ours) | **0.7328** | **0.7285** |

Table 1: **CRATE和基线在测试集上的实验结果。** 计算我们的模型与基线预测跟人工评分之间的皮尔逊相关系数（PCC）和斯皮尔曼等级相关系数（SRCC）。结果表明我们的方法无论基于BERT还是基于RoBERTa，在两个指标上都与人工评分具有最高的相关性。

### 4.3 分析与讨论

#### 4.3.1 消融实验

在本节，我们通过消融实验探讨我们提出的方法中各个组成部分的作用。结果如表2所示。首先，当我们移除排名一致性模块时，模型无法获得精确的译文质量排名信息，导致在两个评

[9] https://github.com/ymcui/Chinese-BERT-wwm
[10] https://huggingface.co/hellonlp/simcse-roberta-base-zh


(c) 2024 中国中文信息学会计算语言学专业委员会

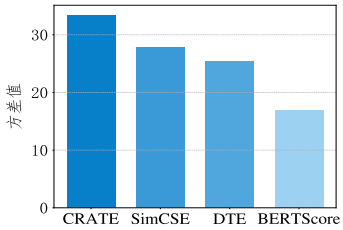



Figure 3: **锚点与负样本之间欧氏距离的方差值。**我们提出的方法计算得到的方差值最大，证明其确实可以捕获负样本之间细微的差别。

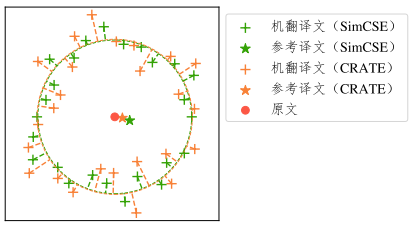



Figure 4: **锚点与负样本之间欧氏距离的案例可视化。**圆圈半径为平均欧式距离，通过样本点到圆周的垂直线长度可以看出，CRATE的样本分布比SimCSE更分散。

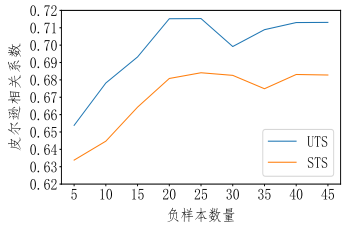



Figure 5: **探究负样本数量对于模型性能的影响。**我们对UTS和STS测试集在不同负样本数量下进行了模型预测与人工评分之间的皮尔逊相关系数计算。

估指标上的结果有所下降。其次，移除对比学习后，模型失去了区分正负样本之间差异的能力，从而导致性能显著降低。另外，当我们仅使用匹配模型或语义相似度模型都会导致性能下降，两者结合使用时效果最佳。再者，当我们移除破坏并重构模块时，性能显著下降。这是因为机翻译文跟参考译文之间的质量差距通常不大，使用它们作为对比学习的正负样本时，模型可能学习到不准确的信息。最后，移除词权重注意力层后，性能略有下降，这表明它在我们的建模中的有效性。值得注意的是，当我们训练匹配模型时不使用原文作为负样本，结果并没有什么变化，这是因为测试集中的译文大多翻译完全，并没有出现漏译的情况，我们在4.3.3节进行了详细的分析。

| 方法 | PCC | SRCC |
|---|---|---|
| CRATE | 0.7153 | 0.7297 |
| w/o 排名一致性 | 0.6924 | 0.6985 |
| w/o 对比学习 | 0.6379 | 0.6149 |
| w/o 匹配度&排名一致性 | 0.6785 | 0.6773 |
| w/o 相似度&排名一致性 | 0.6833 | 0.6917 |
| w/o 原文作为负样本 | 0.7149 | 0.7275 |
| w/o 破坏并重构 | 0.6518 | 0.6397 |
| w/o 词权重注意力层 | 0.7096 | 0.7110 |

Table 2: **我们的模型基于$\mathbf{BERT_{base}}$的消融实验结果。**包含所有模块的模型具有最佳的性能，这证明了我们模型的每个组件的必要性和有效性。w/o 代表删除。

### 4.3.2 译文质量的细粒度排名分析

4.2节的实验结果表明，与其他基准模型相比，我们提出的方法与人工评分的相关性更高。为了探讨我们方法的卓越表现是否源于其区分质量接近的译文的细微差别的能力，我们计算了机翻译文（负样本）与原文（锚点）之间的欧氏距离，并通过分析其方差来进行评估。具体而言，我们对测试集的数据生成一些负样本，并计算平均方差。方差值越大，说明区分负样本细微差异的能力越强。

结果如图3所示，我们的方法（CRATE）计算得到的负样本的方差值最大，这表明我们的方法捕捉到的负样本的质量分布更为分散，证明其确实可以捕捉到质量接近的译文之间的细微差别，从而得到准确的细粒度译文质量排名信息。相比之下，SimCSE方法虽然通过对比学习有效地区分了正样本和负样本，但其在负样本间的细微差异识别上表现较为有限，这可以从其较小的负样本间方差中观察到。此外，DTE在性能上与SimCSE相近，我们认为虽然DTE并未通过对比学习拉开样本之间的距离，但是其通过结合BLEU和语义相似度的优点弥补了这方面的不足。BERTScore计算的方差值最小，这是因为它在处理细粒度质量评估方面的能力较弱，无法有效区分质量不同的负样本。

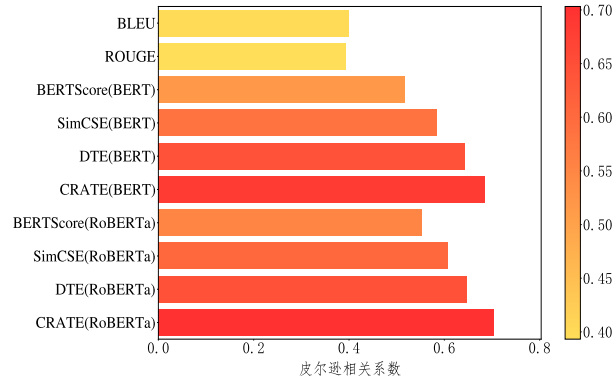



Figure 6: **CRATE和基线在存在漏译情况的测试集（STS）上的实验结果。**实验结果与UTS基本一致。基于预训练模型的评估方法明显优于基于字符串的方法，且相同的指标基于RoBERTa的性能要略优于基于BERT的性能。

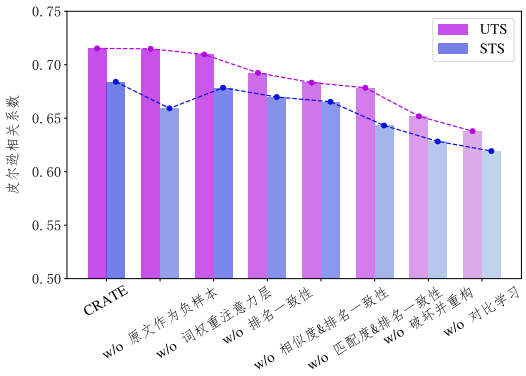



Figure 7: **CRATE在UTS和STS两个测试集上的消融实验对比。**删除“原文作为负样本”后，在几乎未存在漏译的UTS测试集上性能基本不变，在存在漏译的STS测试集上的性能明显下降。

另外，在图4中我们展示了一个具体案例。在对比分析中，我们选用SimCSE作为基准，并在图中呈现了两种方法的性能，其中圆的半径为机翻译文与原文之间的平均距离。观察SimCSE的结果，我们可以发现机翻译文的分布大多集中在圆周附近，这说明SimCSE在区分质量相近的译文方面确实存在不足。相比之下，我们的CRATE方法则呈现出更为分散的分布，表明它能有效捕捉负样本之间的细微差异。综上所述，这些发现突显了我们方法在细粒度译文质量评估方面的有效性。

### 4.3.3 古文漏译情况分析

为了验证模型处理存在漏译现象的译文的能力，我们另外构建了一个特殊的测试集（STS），其中的机翻译文均存在漏译现象。具体做法如下，我们从语料库中随机抽取了4,000条数据对，使用百度翻译将古代汉语翻译成现代汉语，并从中筛选出500条存在漏译的翻译。之后，我们根据原文和参考译文，对这些机翻译文进行了人工评分。结果如图6所示，在STS测试集上的结果与UTS测试集大概一致，基于预训练模型的评估指标明显优于基于字符串的性能指标，且相同指标基于RoBERTa的性能要略优于基于BERT的性能。我们在STS测试集上进行了消融实验分析（见图7），发现其性能走势与UTS测试集基本一致。值得注意的是，当移除原文作为负样本时，在UTS测试集上性能几乎不变，而在STS测试集上则有显著下降。这一结果强调了我们提出的在使用对比学习方法训练原文与译文的匹配度模型时将原文自身作为负样本的方法的有效性。

### 4.3.4 对比学习中超参数的影响

我们在UTS和STS两个测试集上探究了负样本数量对于模型性能的具体影响。如图5所示，在负样本数量增加的初始阶段，模型性能有明显的提升。对于UTS，当负样本数量增至20后，模型的性能提升趋缓，并且在数量增至30时性能略有下降，之后性能回升并保持相对平稳。对STS测试集的观察结果显示，模型性能在负样本数量增至25时达到较高值，增至35时略有回落，随后也表现出较为平稳的趋势。我们认为我们通过破坏并重构的方法生成了更难的负样本，所以才让我们能够用小批量的负样本获得如此的效果。这些结果表明，适量增加负样本的数量可以有效提升模型性能，但超出一定的阈值后，模型性能提升会停滞。基于这些观察，我们将负样本的数量设定为25，以达到模型性能和计算效率的最优平衡。另外，根据经验，我们将对比学习的温度系数设置为0.3。

## 4.4 样例分析

在本节中，我们提供了一些古代汉语翻译质量评估的具体样例，并将我们的方法与SimCSE和BERTScore两个基线进行比较。

在图8中，我们展示了对不同质量的机翻译文的细粒度评估的样例。其中，机翻译文a和b的质量较高，机器译文c、d以及e的质量较低。之前的结果（见图3）表明，SimCSE在译文质量的


（c）2024 中国中文信息学会计算语言学专业委员会

| **古代汉语**：又讨骏之时，殿下在外，实所不综。<br>**参考译文**：另外，声讨杨骏的时候，殿下你还在朝廷之外，委实不曾参与。 | SimCSE | CRATE | Label (Rank) |
|---|---|---|---|
| **机翻译文a**：另外，殿下没有参与谴责杨骏，他那时候在朝廷外。 | 0.8115(2) | 0.8346(1) | 1 |
| **机翻译文b**：声讨杨骏时，殿下你在朝廷外，委实是没有办法。 | 0.8162(1) | 0.8159(2) | 2 |
| **机翻译文c**：我们在责备骏的时候，你在外，实在是不了解。 | 0.5516(4) | 0.5647(3) | 3 |
| **机翻译文d**：在讨论杨骏的时候，你还在外面，实在是过分。 | 0.5405(5) | 0.5328(4) | 4 |
| **机翻译文e**：你在夸奖杨骏的时候，确实并没有苦衷。 | 0.5584(3) | 0.4935(5) | 5 |

Figure 8: **不同质量的机翻译文的细粒度评估结果样例。**结果包含了评分和排名，其中的标签由人工进行标注。结果表明，SimCSE得到的细粒度排名并不准确，而我们的模型CRATE对译文质量的细粒度排名与标签完全一致。

| 样例 | BERTScore | CRATE | Label |
|---|---|---|---|
| **古代汉语**：尊兄**应期赞世**，**配业光国**，魄兆见矣。<br>**参考译文**：尊兄您顺应天时，把握时机，辅佐光大国家事业，智慧的光芒已经显现。<br>**机翻译文**：哥哥**应期赞世**人，**配业光国**，魄预见了。 | 0.7648 | 0.3573 | 0.2 |
| **古代汉语**：今以贵人为皇后，**使行丞相事左将军向朗持节授玺绶**。<br>**参考译文**：现在封贵人为皇后，派代理丞相事宜的左将军向郎持符节授予印信绶带。<br>**机翻译文**：现在用贵人为皇后，**使行丞相事左将军向朗持节授玺绶**。 | 0.8849 | 0.5409 | 0.6 |

Figure 9: **存在漏译情况的译文评估结果的两组样例。**红色字体标记了机翻译文中漏译的部分以及其在古代汉语中的相应位置，标签是人工评分。结果表明，BERTScore给存在漏译的译文生成了偏高的分数，而我们的模型评分结果与人工评分更接近。

细粒度分析方面比DTE表现更优，因此我们将SimCSE作为了基准方法。通过SimCSE的评估结果，我们观察到a和b的得分显著高于c、d和e的得分，这证实了SimCSE在区分质量差距较大的译文方面的有效性。然而，其对译文a和b之间，以及c、d和e之间的排名与实际标签并不一致，这暴露了它在获取精确细粒度排名上的局限性。相反，我们提出的模型不仅明确区分了高质量和低质量的译文，而且得到的细粒度的译文质量排名信息与标签完全一致，显著展现了其在捕捉质量接近译文间微妙差异的能力。

在图9中，我们展示了两组存在漏译情况的译文的评估样例，并用红色字体高亮显示了机翻译文中漏译的部分以及其在古代汉语中的相应位置。评估结果表明，BERTScore指标明显高于人工评分。相比之下，我们模型的预测结果与人工评分更为一致，这表明我们的方法能够为存在漏译现象的译文提供更准确的评分，进一步证实了我们提出的方法的有效性和优越性。

## 5 结论

在本文中，我们提出了一种基于对比学习和排名一致性的古代汉语翻译评估模型（CRATE）。通过确保机翻译文在与参考译文语义相似度上的排名以及与古代汉语匹配度上的排名之间的一致性，CRATE捕获到了细粒度的译文质量排名信息。使用对比学习训练原文和译文的匹配模型时，将原文自身作为了负样本，给存在漏译情况的译文更合理的分数。在测试集上的实验结果表明，CRATE的性能超过了现有的评估方法，与人类评分具有更高的相关性。大量的实验分析也证明了我们模型中每个组成部分的有效性和合理性。尽管预训练模型表现出较好的性能，但在古文的语义理解方面仍有提升空间。在未来的研究中，我们将探讨如何利用大语言模型进行评估。然而，目前大语言模型在古文方面的训练语料极为有限，这也是我们亟需关注的重点。

## 致谢

本成果受国家自然科学基金项目（62306045，61872402），北京语言大学校级项目（中央高校基本科研业务费专项资金）（18ZDJ03），北京语言大学梧桐创新平台项目（21PT04），


(c) 2024 中国中文信息学会计算语言学专业委员会

北京语言大学研究生创新基金项目（中央高校基本科研业务费专项资金）（24YCX085）资助。

## 参考文献

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare R. Voss, editors, *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, pages 65–72. Association for Computational Linguistics.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to rank: from pairwise approach to listwise approach. In Zoubin Ghahramani, editor, *Machine Learning, Proceedings of the Twenty-Fourth International Conference (ICML 2007), Corvallis, Oregon, USA, June 20-24, 2007*, volume 227 of *ACM International Conference Proceeding Series*, pages 129–136. ACM.

Kenneth Ward Church and Eduard H. Hovy. 1993. Good applications for crummy machine translation. *Mach. Transl.*, 8(4):239–258.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6894–6910. Association for Computational Linguistics.

Hui Huang, Shuangzhi Wu, Kehai Chen, Hui Di, Muyun Yang, and Tiejun Zhao. 2023. Improving translation quality estimation with bias mitigation. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 2175–2190. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Juntao Li, Yan Song, Haisong Zhang, Dongmin Chen, Shuming Shi, Dongyan Zhao, and Rui Yan. 2018. Generating classical chinese poems via conditional variational autoencoder and adversarial training. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3890–3900. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Dayiheng Liu, Kexin Yang, Qian Qu, and Jiancheng Lv. 2020. Ancient-modern chinese translation with a new large training dataset. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 19(1):6:1–6:13.

Jiduan Liu, Jiahao Liu, Qifan Wang, Jingang Wang, Wei Wu, Yunsen Xian, Dongyan Zhao, Kai Chen, and Rui Yan. 2023. Rankcse: Unsupervised sentence representations learning via learning to rank. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13785–13802. Association for Computational Linguistics.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (mqm): A framework for declaring and describing translation quality metrics. *Tradumàtica*, (12):0455–463.


(c) 2024 中国中文信息学会计算语言学专业委员会

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Tharindu Ranasinghe, Constantin Orasan, and Ruslan Mitkov. 2020. Transquest: Translation quality estimation with cross-lingual transformers. In Donia Scott, Núria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 5070–5081. International Committee on Computational Linguistics.

Matthew G. Snover, Bonnie J. Dorr, Richard M. Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers, AMTA 2006, Cambridge, Massachusetts, USA, August 8-12, 2006*, pages 223–231. Association for Machine Translation in the Americas.

Keh-Yih Su, Ming-Wen Wu, and Jing-Shin Chang. 1992. A new quantitative quality measure for machine translation systems. In *14th International Conference on Computational Linguistics, COLING 1992, Nantes, France, August 23-28, 1992*, pages 433–439.

Christoph Tillmann, Stephan Vogel, Hermann Ney, Arkaitz Zubiaga, and Hassan Sawaf. 1997. Accelerated DP based search for statistical translation. In George Kokkinakis, Nikos Fakotakis, and Evangelos Dermatas, editors, *Fifth European Conference on Speech Communication and Technology, EUROSPEECH 1997, Rhodes, Greece, September 22-25, 1997*, pages 2667–2670. ISCA.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Kexin Yang, Dayiheng Liu, Qian Qu, Yongsheng Sang, and Jiancheng Lv. 2021. An automatic evaluation metric for ancient-modern chinese translation. *Neural Comput. Appl.*, 33(8):3855–3867.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Siqing Zhou and Shijing Si. 2023. Evaluating the capability of chatgpt on ancient chinese. *CoRR*, abs/2312.15304.

CCL 2024

# 基于两种新颖辅助任务的端到端语音翻译

**窦怀厦，吕孟哲，李军辉**
苏州大学计算机科学与技术学院，江苏省苏州市
{20225227069,20225227080}@stu.suda.edu.cn，lijunhui@suda.edu.cn

## 摘要

端到端语音翻译具有跨模态和跨语言的特性，如何有效地利用这些特性是一个具有挑战性的问题。本文基于多任务学习框架，提出两种新颖辅助任务。语音增强的文本翻译任务通过在文本翻译任务中融入语音模态信息来缓解语音和文本的模态差异，最终提升语音翻译任务的性能。全局感知条件掩码语言建模任务能够同时建模转录文本和译文进而利用文本的全局上下文信息指导翻译模型的训练。在MuST-C数据集8个语向的实验结果表明，本文的方法显著优于基线系统，并且达到了与其它端到端语音翻译系统可竞争的性能水平。进一步的分析实验表明，本文的方法能够缓解语音和文本之间的模态差异并且在不损害文本翻译任务性能的情况下提升语音翻译任务的性能。

**关键词：** 端到端语音翻译；多任务学习；辅助任务

# End-to-End Speech Translation Enhanced by Two Novel Auxiliary Tasks

**Huaixia Dou, Mengzhe Lyu, Junhui Li**
School of Computer Science and Technology, Soochow University, Suzhou, China
{20225227069,20225227080}@stu.suda.edu.cn，lijunhui@suda.edu.cn

## Abstract

End-to-end speech translation is characterized by cross-modal and cross-linguistic properties. How to better utilize these properties is a challenging question. In this paper, we propose two novel auxiliary tasks based on a multi-task learning framework. The speech augmented text translation task alleviates the modality gap between speech and text by incorporating speech modality information into the text translation task, which ultimately improves the performance of the speech translation task. The globally aware conditional mask language modeling task is able to model both the transcription and the translation simultaneously and then use the global context information of the text to guide the training of the translation model. Experimental results on the MuST-C benchmark show that the approach in this paper significantly outperforms the baseline system and achieves a competable performance with other end-to-end speech translation systems. Further analysis shows that our approach is able to alleviate the modality gap between speech and text and preserve the performance of the original machine translation task.

**Keywords:** End-to-end speech translation , Multi-task learning , Auxiliary tasks

第二十三届中国计算语言学大会论文集，第183页-第196页，太原，中国，2024年7月25日至28日。
卷1：主会论文
(c) 2024 中国中文信息学会计算语言学专业委员会

## 1 引言

语音翻译旨在将源语言的语音转化为目标语言的文本。传统的语音翻译系统通常采用级联方法 (Dong et al., 2019; Zhang et al., 2019; Sperber and Paulik, 2020; Lam et al., 2021)，首先使用自动语音识别系统将源语言的语音转录为文本，接着使用机器翻译系统将转录文本翻译为目标语言的文本。由于包含多个中间步骤，级联系统存在错误传播和高延迟的问题。端到端语音翻译因为无需生成中间的转录文本，有望克服级联系统的缺陷，因此在近年来受到越来越多研究者的关注 (Wang et al., 2020b; Ren et al., 2020; Zhang et al., 2020a; Huang et al., 2021; Lam et al., 2022; Han et al., 2023)。目前端到端语音翻译的主流做法是在多任务学习框架下引入多个相关任务来改善语音翻译任务的训练，本文的目标是提出更好的辅助任务来提升语音翻译任务的性能。

在多任务学习框架下，语音翻译任务是主任务，而文本翻译任务通常被用来辅助语音翻译任务的训练 (Ye et al., 2021; Tang et al., 2021; Ye et al., 2022; Yin et al., 2023; Zhang et al., 2023b)。传统的文本翻译任务只涉及单一模态，而语音模态中富含的语义信息可以为文本翻译任务提供额外的上下文和语音特征。因此，本文提出语音增强的文本翻译(Speech Augmented Text Translation, SATT)任务，在文本翻译过程中融合语音模态的信息来缓解语音和文本的模态差异，进而提升语音翻译任务的性能。此外，当前的工作主要集中在同时对跨模态的源端语音和转录文本进行建模，而对同模态的转录文本和译文关注较少。Zhou等人(2022)研究表明，译文中包含的上下文信息能够进一步提升文本翻译任务的性能。因此，本文提出全局感知条件掩码语言建模(Globally Aware Conditional Masked Language Modeling, GACMLM)任务来同时建模转录本文和译文并且利用文本的全局上下文信息来辅助翻译模型的训练。

具体地，针对SATT任务，本文对编码器中的自注意力模块进行修改，使得在自注意力计算过程中，转录文本能够同时关注到文本模态和语音模态的源端上下文信息。针对GACMLM任务，本文借鉴Zhou等人(2022)的思路，采用条件掩码语言模型 (Ghazvininejad et al., 2019)来建模转录文本和译文的双向全局上下文。整个训练过程分为三个阶段。在第一个阶段，利用外部机器翻译语料来预训练模型的翻译编码器和翻译解码器模块；在第二个阶段，利用语音增强的文本翻译任务辅助语音翻译任务的训练，此外在该阶段以多任务学习的方式同时训练条件掩码语言模型；在第三个阶段，将训练好的条件掩码语言模型作为教师模型，通过知识蒸馏 (Hinton et al., 2015)同时辅助语音翻译任务和语音增强的文本翻译任务的训练。在MuST-C八个语向的实验结果表明，本文方法的平均BLEU值能够达到30.4，相比基线系统提升1.5个点。进一步的分析实验表明，本文的方法能够缓解语音和文本的模态差异，同时在不损害文本翻译任务性能的情况下提升语音翻译任务的性能。

## 2 相关工作

### 2.1 端到端语音翻译

为了应对级联系统中固有的错误传播和高时延等问题，端到端语音翻译被提出来直接将源语言的语音映射为目标端文本，而无需生成中间的临时转录文本。作为跨模态和跨语言的任务，端到端语音翻译可获得的数据远少于语音识别和文本翻译的数据。为了应对语音翻译数据不足的问题，Ye等人 (2019)；Bahar等人 (2019)；Pino等人(2019)利用语音识别和文本翻译数据合成了更多的语音翻译数据，Pino等人 (2020)；Wang等人(2021)使用自训练技术进行数据增强，Lam等人 (2022)采用反向翻译和知识蒸馏来获取额外的训练数据。为了应对建模困难问题 (Xu et al., 2023)，多任务学习框架被广泛使用 (Ye et al., 2021; Ye et al., 2022; Fang et al., 2022; Zhou et al., 2023)。在多任务框架下，文本翻译任务通常作为辅助任务，在此基础上，为了缩小语音和文本的模态差异，Han等人 (2021)使用共享的语义记忆模块把语音和文本特征投影到公共的语义空间中，Ye等人 (2022)；Quyang等人 (2023)；Zhang等人 (2022a)采用对比学习方法来拉近语音表示和文本表示的距离，Fang等人 (2022)；Zhou等人 (2023)；Cheng等人 (2023)采用混合语音序列和文本序列的方法来缓解跨模式表征差异。区别于先前的工作，本文在常规的文本翻译任务中利用注意力机制同语音模态进行交互。一方面语音模态中富含的语义信息可以为文本翻译任务提供额外的上下文和语音特征，另一方面转录文本与语音模态的交

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版
**项目资助**：江苏高校优势学科建设工程资助项目

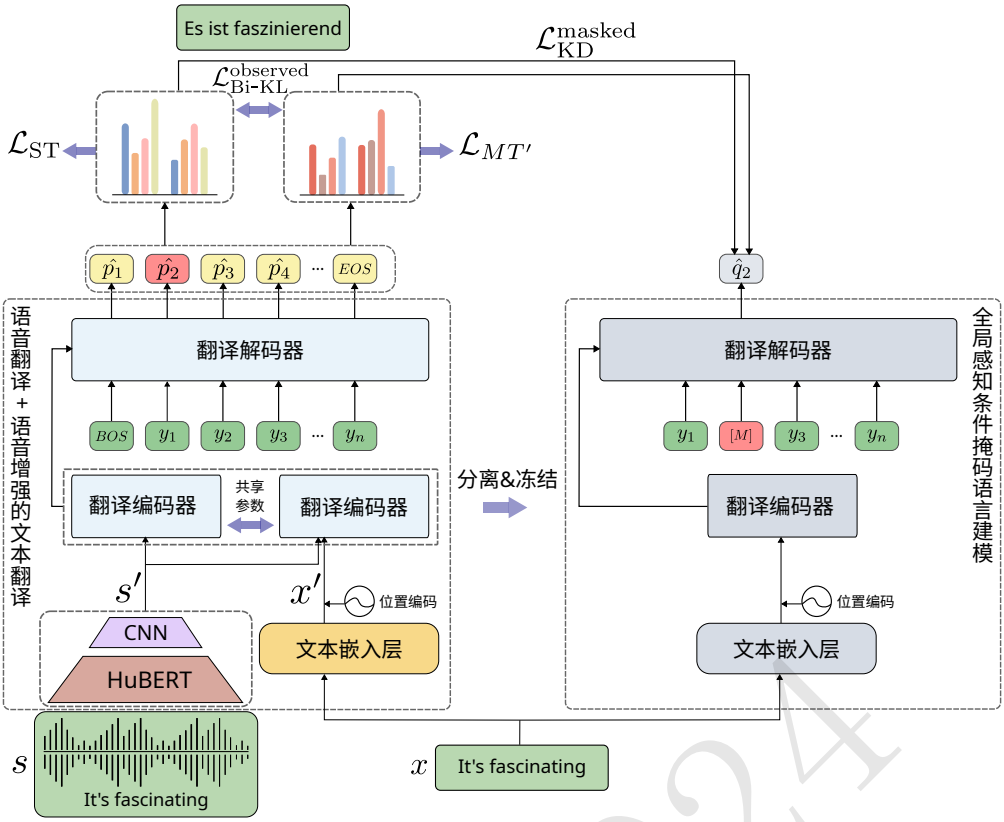



图 1. 本文模型架构(第三阶段)

互能够缩小语音和文本的模态差异。此外，相关工作中很少涉及到建模目标端译文，而译文中包含的上下文信息能够为翻译任务提供额外的帮助。

### 2.2 条件掩码语言模型

条件掩码语言模型通常采用传统的编码器-解码器架构，利用源端输入文本作为条件，目标端可见单词作为上下文来预测被遮盖的单词。相比于神经翻译模型，条件掩码语言模型不仅能够关注源端的文本信息，而且能够关注目标端的上下文信息。Ghazvininejad等人 (2019)使用条件掩码语言模型预测被遮盖的单词，并结合高效的并行迭代解码算法来非自回归地生成译文。Higuchi等人 (2020)组合连接时序分类(CTC)和条件掩码语言模型来改善CTC的预测结果。Chen等人 (2021)将其用于非自回归语音识别任务并且额外增加约束来缩小训练与推理时的差异。Inaguma等人 (2021)将其用于端到端语音翻译的非自回归解码。Chen等人 (2020)；Zhou等人(2022)将其作为教师模型，利用知识蒸馏来改善文本生成任务和文本翻译任务的训练。本文受Zhou等人 (2022)启发，提出全局感知条件掩码语言建模任务来建模文本的全局上下文信息，并且利用这部分信息来辅助翻译模型的训练。

## 3 方法

本文在多任务学习框架下提出了两个新颖辅助任务，即语音增强的文本翻译任务和全局感知条件掩码语言建模任务，旨在改善语音翻译任务的训练。本节首先对端到端语音翻译进行建模，接着依次介绍这几个辅助任务。

### 3.1 问题建模

语音翻译的训练数据通常由包含源语言语音、源语言转录文本和目标语言译文的三元组构成，形式化表示为$\mathcal{D}_{\text{ST}} = \{(\mathbf{s}, \mathbf{x}, \mathbf{y})\}$。其中$\mathbf{s} = (s_1, s_2, ..., s_{|\mathbf{s}|})$是语音帧序列，$\mathbf{x} = (x_1, x_2, ..., x_{|\mathbf{x}|})$是转录文本序列，$\mathbf{y} = (y_1, y_2, ..., y_{|\mathbf{y}|})$是译文序列，$|\mathbf{s}|$、$|\mathbf{x}|$和$|\mathbf{y}|$分别是语音帧序列、转录文本序列和译文序列的长度。端到端语音翻译系统需要直接将语音帧序列$\mathbf{s}$转换为译


(c) 2024 中国中文信息学会计算语言学专业委员会

文序列$\mathbf{y}$而无需生成中间的转录文本序列$\mathbf{x}$。除此之外，本文还利用外部机器翻译数据对模型进行预训练，这部分数据记为$\mathcal{D}_{\mathrm{EXT}}=\{(\mathbf{x}^{\mathbf{ext}},\mathbf{y}^{\mathbf{ext}})\}$。

### 3.2 语音翻译

本文使用声学编码器来提取语音特征，利用翻译编码器对语音特征进一步编码，最终通过翻译解码器来完成目标译文的解码。

#### 3.2.1 编码过程

预训练的HuBERT (Hsu et al., 2021)模型和两个额外的卷积层作为声学编码器来提取语音$\mathbf{s}$的底层特征。原始波形信号的采样频率为16kHz，经过HuBERT模型后得到语音特征序列。由于语音特征序列的长度通常远大于其转录文本的长度，为了进一步匹配语音特征序列和转录文本的长度，本文在HuBERT之后使用两个卷积层来缩短语音特征序列的长度。声学编码器和卷积层对语音$\mathbf{s}$的编码过程如公式(1)所示：

$$\mathbf{s}'=\mathrm{CNN}\left(\mathrm{HuBERT}\left(\mathbf{s}\right)\right), \tag{1}$$

其中，HuBERT(·)表示使用HuBERT模型进行编码，CNN(·)表示卷积操作。

翻译编码器用于进一步对声学编码器的输出$\mathbf{s}'$编码，翻译编码器在结构上与标准的Transformer (Vaswani et al., 2017)编码器相同，由$N_e$个完全相同的层堆叠而成，每层由一个多头掩码自注意力子层和一个全连接前馈神经网络构成。在语音翻译任务中，多头掩码自注意力子层使用前一层的隐状态作为输入并通过多头注意力模块和层规范化模块进行计算，如公式(2)所示：

$$B^{(i)}=\mathrm{LayerNorm}\left(\mathrm{MultiHead}\left(A^{(i-1)},A^{(i-1)},A^{(i-1)}\right)+A^{(i-1)}\right), \tag{2}$$

其中，$1<=i<=N_e$，$A^{(i-1)}\in\mathbb{R}^{t_s\times d_{\mathrm{model}}}$表示第$i-1$层的隐状态，$t_s$为经过卷积层后的语音特征序列的长度，$d_{\mathrm{model}}$为模型维度，$A^{(0)}=\mathbf{s}'$为声学编码器的输出，MultiHead(·)表示多头注意力模块，LayerNorm(·)表示层规范化，$B^{(i)}\in\mathbb{R}^{t_s\times d_{\mathrm{model}}}$。接下来，全连接前馈神经网络使用$B^{(i)}$作为输入并经过层规范化模块计算得到当前层的语音编码输出$A^{(i)}$，如公式(3)所示：

$$A^{(i)}=\mathrm{LayerNorm}\left(\mathrm{FeedForward}\left(B^{(i)}\right)+B^{(i)}\right), \tag{3}$$

其中，LayerNorm(·)表示层规范化，FeedForward(·)表示全连接前馈神经网络。为了简便起见，本文定义$\mathbf{a}=A^{(N_e)}$表示语音经过翻译编码器后得到的上下文表示。

#### 3.2.2 解码过程

翻译解码器使用标准的Transformer解码器，由$N_d$个完全相同的层堆叠而成。每层由一个多头掩码自注意力子层，一个多头上下文注意力子层和一个全连接前馈神经网络构成。本文使用交叉熵损失作为语音翻译任务的损失函数，如公式(4)所示：

$$\mathcal{L}_{\mathrm{ST}}(\mathbf{s},\mathbf{y})=-\sum_{i=1}^{|\mathbf{y}|}\log P(\mathbf{y}_i|\mathbf{a},\mathbf{y}_{<i}). \tag{4}$$

### 3.3 语音增强的文本翻译

为了缓解模态差异，本文提出语音增强的文本翻译任务来融入语音模态的信息。在多任务学习框架下，语音增强的文本翻译任务和语音翻译任务共享翻译编码器与翻译解码器。

#### 3.3.1 编码过程

文本词嵌入层和位置编码模块用来编码转录文本$\mathbf{x}$，如公式(5)所示：

$$\mathbf{x}'=\mathrm{Embedding}\left(\mathbf{x}\right)+\mathrm{Pos}\left(\mathbf{x}\right), \tag{5}$$

其中，Embedding(·)表示文本词嵌入层，Pos(·)表示位置编码模块。


(c) 2024 中国中文信息学会计算语言学专业委员会

为了有效地利用语音模态中富含的语义信息，本文在对转录文本进行多头自注意力计算时，通过将语音和转录文本的键值对向量进行拼接来融合语音模态的信息。具体地，声学编码器的输出$\mathbf{s}'$将和翻译编码器中每一层的隐状态输出$A^{(i-1)}$进行拼接作为新的文本键值对向量参与多头注意力计算，如公式(6)和公式(7)所示：

$$Q^{(i-1)} = A^{(i-1)}, K^{(i-1)} = V^{(i-1)} = \text{Concat}(\mathbf{s}', A^{(i-1)}), \tag{6}$$

$$B^{(i)} = \text{LayerNorm}(\text{MultiHead}(Q^{(i-1)}, K^{(i-1)}, V^{(i-1)}) + A^{(i-1)}), \tag{7}$$

其中，$1 <= i <= N_e$，$A^{(0)} = \mathbf{x}'$。接下来利用公式(3)，以$B^{(i)}$作为输入并经过层规范化模块计算得到当前层的文本编码输出$A^{(i)}$。为了简便起见，本文定义$\mathbf{e_{mix}} = A^{(N_e)}$表示文本经过翻译编码器并且进行语音增强后得到的上下文表示。

### 3.3.2 解码过程与模态正则化

翻译解码器以$\mathbf{e_{mix}}$作为输入，通过交叉熵损失函数来计算语音增强的文本翻译任务的损失，如公式(8)所示：

$$\mathcal{L}_{\text{MT}'}(\mathbf{s}, \mathbf{x}, \mathbf{y}) = -\sum_{i=1}^{|\mathbf{y}|} \log P(\mathbf{y}_i | \mathbf{e_{mix}}, \mathbf{y}_{<i}). \tag{8}$$

为了缓解语音和文本的模态差异，本文希望模型分别以语音上下文表示$\mathbf{a}$和以语音增强后的文本上下文表示$\mathbf{e_{mix}}$作为条件进行预测的目标端概率分布趋于一致。本文使用双向KL散度作为模态正则化损失来衡量两个预测概率的不一致性程度，如公式(9)所示：

$$\begin{aligned}\mathcal{L}^{\text{all}}_{\text{Bi-KL}}(\mathbf{s}, \mathbf{x}, \mathbf{y}) = \sum_{i=1}^{|\mathbf{y}|} \frac{1}{2} (&\mathcal{D}_{\text{KL}}(P(\mathbf{y}_i | \mathbf{a}, \mathbf{y}_{<i}) || P(\mathbf{y}_i | \mathbf{e_{mix}}, \mathbf{y}_{<i})) \\ &+ \mathcal{D}_{\text{KL}}(P(\mathbf{y}_i | \mathbf{e_{mix}}, \mathbf{y}_{<i}) || P(\mathbf{y}_i | \mathbf{a}.\mathbf{y}_{<i}))).\end{aligned} \tag{9}$$

## 3.4 全局感知条件掩码语言建模

为了从文本模态中学习更多的上下文信息，本文提出全局感知条件掩码语言建模任务来同时建模源端的转录文本和目标端的译文。该任务可分为文本全局上下文学习与知识蒸馏两个子任务。文本全局上下文学习任务通过训练条件掩码语言模型来学习文本的全局上下文信息。知识蒸馏任务通过分离出训练好的条件掩码语言模型并使用该模型进一步指导语音翻译任务和语音增强的文本翻译任务的训练。

### 3.4.1 文本全局上下文学习

条件掩码语言模型按照概率$p$随机对目标端译文执行掩码操作。目标端译文$\mathbf{y}$中的单词将被划分为两个集合。其中，掩码单词集合记为$\mathbf{y}_{masked}$，可见单词集合记为$\mathbf{y}_{observed}$。针对掩码单词集合$\mathbf{y}_{masked}$，其在原句中的对应位置被替换为特殊的掩码标记$[M]$。条件掩码语言模型会根据转录文本$\mathbf{x}$和目标端可见单词集合$\mathbf{y}_{observed}$来预测目标端掩码单词集合$\mathbf{y}_{masked}$。

与Zhou等人(2022)不同，本文的条件掩码语言模型和语音翻译任务共享文本词嵌入层、翻译编码器与翻译解码器。条件掩码语言模型首先使用文本词嵌入层和位置编码模块来对转录文本$\mathbf{x}$进行编码得到$\mathbf{x}'$，接着翻译编码器对$\mathbf{x}'$进一步编码得到文本的上下文表示$\mathbf{e}$。为了使翻译解码器适应条件掩码语言建模任务，本文移除了翻译解码器中多头掩码自注意力子层的注意力掩码，因此条件掩码语言模型能够同时关注到转录文本、译文的历史信息和译文的未来信息，它应该包含双向的文本全局上下文信息。本文使用交叉熵损失作为全局感知条件掩码语言建模任务的损失函数，如公式(10)所示：

$$\mathcal{L}_{\text{CMLM}}(\mathbf{x}, \mathbf{y}) = -\sum_{\mathbf{y}_i \in \mathbf{y}_{masked}} \log P(\mathbf{y}_i | \mathbf{e}, \mathbf{y}_{observed}). \tag{10}$$

CCL 2024

#### 3.4.2 知识蒸馏

在文本全局上下文学习完成后，本文从文本词嵌入层、翻译编码器和翻译解码器中分离出训练好的条件掩码语言模型。如图1右侧所示，分离出的条件掩码语言模型的参数不再更新，此时把条件掩码语言作为教师模型，利用知识蒸馏来进一步辅助左侧翻译模型的训练。针对掩码单词集合$\mathbf{y}_{masked}$，教师模型利用学习到的文本全局上下文信息来进一步改善语音翻译任务和语音增强的文本翻译任务的训练。如公式(11)所示：

$$\begin{aligned}\mathcal{L}_{\text{KD}}^{\text{masked}}(\mathbf{s},\mathbf{x},\mathbf{y}) = \sum_{\mathbf{y}_i \in \mathbf{y}_{masked}} & (\mathcal{D}_{\text{KL}}(P(\mathbf{y}_i|\mathbf{e},\mathbf{y}_{observed})||P(\mathbf{y}_i|\mathbf{e}_{\mathbf{mix}},\mathbf{y}_{<i})) \\ & +\mathcal{D}_{\text{KL}}(P(\mathbf{y}_i|\mathbf{e},\mathbf{y}_{observed})||P(\mathbf{y}_i|\mathbf{a},\mathbf{y}_{<i}))).\end{aligned} \tag{11}$$

### 3.5 训练方式

整个训练过程分为三个阶段。在第一个阶段，采用外部机器翻译语料对模型的翻译编码器和翻译解码器进行训练。该阶段的损失函数如公式(12)所示：

$$\mathcal{L}_{\text{stage1}}(\mathbf{x}^{\mathbf{ext}},\mathbf{y}^{\mathbf{ext}}) = -\sum_{i=1}^{|\mathbf{y}^{\mathbf{ext}}|} \log P(\mathbf{y}_i^{\mathbf{ext}}|\mathbf{x}^{\mathbf{ext}},\mathbf{y}_{<i}^{\mathbf{ext}}). \tag{12}$$

在第二个阶段，通过多任务学习来联合训练语音翻译任务、语音增强的文本翻译任务和文本全局上下文学习任务。该阶段的损失函数如公式(13)所示：

$$\mathcal{L}_{\text{stage2}}(\mathbf{s},\mathbf{x},\mathbf{y}) = \mathcal{L}_{\text{ST}}(\mathbf{s},\mathbf{y}) + \mathcal{L}_{\text{MT}'}(\mathbf{s},\mathbf{x},\mathbf{y}) + \mathcal{L}_{\text{Bi-KL}}^{\text{all}}(\mathbf{s},\mathbf{x},\mathbf{y}) + \mathcal{L}_{\text{CMLM}}(\mathbf{x},\mathbf{y}). \tag{13}$$

在第三个阶段，对于可见单词集合$\mathbf{y}_{observed}$，继续使用双向KL散度约束语音翻译任务和语音增强的文本翻译任务的目标端预测的概率分布。如公式(14)所示：

$$\begin{aligned}\mathcal{L}_{\text{Bi-KL}}^{\text{observed}}(\mathbf{s},\mathbf{x},\mathbf{y}) = \sum_{\mathbf{y}_i \in \mathbf{y}_{observed}} & \frac{1}{2}(\mathcal{D}_{\text{KL}}(P(\mathbf{y}_i|\mathbf{a},\mathbf{y}_{<i})||P(\mathbf{y}_i|\mathbf{e}_{\mathbf{mix}},\mathbf{y}_{<i})) \\ & +\mathcal{D}_{\text{KL}}(P(\mathbf{y}_i|\mathbf{e}_{\mathbf{mix}},\mathbf{y}_{<i})||P(\mathbf{y}_i|\mathbf{a},\mathbf{y}_{<i}))),\end{aligned} \tag{14}$$

对于掩码单词集合$\mathbf{y}_{masked}$，利用知识蒸馏把训练好的条件掩码语言模型作为教师模型，进一步提升语音翻译任务和语音增强的文本翻译任务的性能。该阶段的损失函数如公式(15)所示：

$$\mathcal{L}_{\text{stage3}}(\mathbf{s},\mathbf{x},\mathbf{y}) = \mathcal{L}_{\text{ST}}(\mathbf{s},\mathbf{y}) + \mathcal{L}_{\text{MT}'}(\mathbf{s},\mathbf{x},\mathbf{y}) + \mathcal{L}_{\text{Bi-KL}}^{\text{observed}}(\mathbf{s},\mathbf{x},\mathbf{y}) + \alpha\mathcal{L}_{\text{KD}}^{\text{masked}}(\mathbf{s},\mathbf{x},\mathbf{y}), \tag{15}$$

其中，$\alpha$是超参数，控制知识蒸馏任务的权重[0]。

## 4 实验

### 4.1 数据集

**语音翻译数据集** MuST-C (Di Gangi et al., 2019)是一个多语言的语音翻译数据集。参考相关的研究工作，本文在英语(EN)到其它八种语言的翻译方向上进行实验，这八种语言包括德语(DE)、法语(FR)、俄语(RU)、西班牙语(ES)、意大利语(IT)、罗马尼亚语(RO)、葡萄牙语(PT)和荷兰语(NL)。每个翻译方向都包含至少385个小时的录音数据。本文使用验证集对模型进行验证，并且在测试集tst-COMMON上报告最终结果。

**机器翻译数据集** 本文使用的外部机器翻译数据集包括用于英语-德语/法语/俄语/西班牙语/罗马尼亚语方向的WMT数据集和用于英语-意大利语/葡萄牙语/荷兰语的OPUS100数据集[1] (Zhang et al., 2020b)。详细的统计数据可参见表1，其中列出了语音翻译数据集和外部机器翻译数据集的相关信息。

**预处理** 本文使用原始格式的16位16kHz单声道音频作为语音输入。转录文本和译文均区分大小写。对于每个翻译方向，本文利用单字元(Unigram)SentencePiece[2]模型在语音翻译数据集上进行训练来学习转录文本和译文间的共享词表。接下来，使用共享词表把语音翻译语料和外部机器翻译语料分割成子词单元。本文将共享词表大小统一设置为10K。

[0]考虑到第三阶段$\mathcal{L}_{\text{Bi-KL}}^{\text{observed}}$与第二阶段$\mathcal{L}_{\text{Bi-KL}}^{\text{all}}$的区别仅仅在于译文$\mathbf{y}$中选中的单词范围不同，为了保持一致性，本文没有将其包含在$\alpha$的控制范围

[1]http://opus.nlpl.eu/opus-100.php

[2]https://github.com/google/sentencepiece


(c) 2024 中国中文信息学会计算语言学专业委员会

| 英语 (EN)→ | MuST-C | | 外部机器翻译数据 | |
|---|---|---|---|---|
| | 小时数 | 句子数 | 来源 | 句子数 |
| **德语** (DE) | 408 | 234K | WMT16 | 4.6M |
| **法语** (FR) | 492 | 280K | WMT16 | 40.8M |
| **俄语** (RU) | 489 | 270K | WMT16 | 2.5M |
| **荷兰语** (NL) | 442 | 253K | OPUS100 | 1.0M |
| **西班牙语** (ES) | 504 | 270K | WMT13 | 15.2M |
| **意大利语** (IT) | 465 | 258K | OPUS100 | 1.0M |
| **葡萄牙语** (PT) | 385 | 211K | OPUS100 | 1.0M |
| **罗马尼亚语** (RO) | 432 | 240K | WMT16 | 0.6M |

表 1. 实验数据集统计

### 4.2 实验设置

**模型设置** 本文使用预训练的HuBERT模型来提取语音特征。在HuBERT[3]之后添加了两层卷积层来进一步缩短语音特征序列的长度，卷积层内核大小为5，步长为2，填充大小为2，隐层维度为512。本文翻译编码器和翻译解码器层数$N_e$和$N_d$均设置为6，多头注意力设置为8个头，隐层维度设置为512、全连接前馈神经网络中间层维度设置为2048。

**训练与推理** 本文使用Adam优化器 (Kingma and Ba, 2015)，其中，$\beta_1 = 0.9, \beta_2 = 0.98$。在整个训练过程中，文本批处理大小设置为33K个令牌以内，语音批处理大小设置为16M语音帧以内，预热(WarmUp)步数设置为4K，Dropout和标签平滑(Label Smoothing)设置为0.1。第一阶段的学习率设置为7e-4，后续阶段的学习率设置为1e-4。为了避免模型过拟合，在第一阶段和第三阶段，当验证集上的BLEU值超过10轮不再提升则停止该阶段训练；在第二阶段，当验证集上的条件掩码语言模型损失值$\mathcal{L}_{\text{CMLM}}$超过5轮不再降低则停止该阶段训练。本文中知识蒸馏权重$\alpha$和掩码概率$p$最终分别设置为0.5和0.15。与训练过程不同，推理过程中不再使用转录文本和译文信息，因此本文仅保留语音翻译模块进行推理。为了公平比较，本文平均最后10个检查点用于评估模型性能。本文使用集束搜索(Beam Search)并且设置Beam Size为8。针对八个翻译方向，本文设置了不同的长度惩罚值，分别是1.1(英语-德语)、1.7(英语-法语)、0.4(英语-俄语)、0.5(英语-西班牙语)、1.6(英语-罗马尼亚语)、0.7(英语-意大利语)、1.4(英语-葡萄牙语)、0.9(英语-荷兰语)。本文的模型使用FairSeq[4] (Ott et al., 2019)实现，并且在4张Nvidia Tesla V100上训练得到。

**模型评估** 对于所有翻译方向，本文使用SacreBLEU[5] (Post, 2018)来报告大小写敏感并且去标记化(Detokenized)后的BLEU分数 (Papineni et al., 2002)。

### 4.3 实验结果

为了公平起见，本文方法将同多个端到端语音翻译模型进行比较，这些模型均使用了外部机器翻译数据。我们根据模型使用的声学编码器分为两类：基于Wav2vec2.0 (Baevski et al., 2020)的模型和基于HuBERT的模型。

- 基于Wav2Vec2.0的模型：Chimera (Han et al., 2021)通过将语音和文本表征投影到共同的语义表征空间中来缩小模态差异；XSTNet (Ye et al., 2021)提出了一个多任务训练框架，并采用渐进式训练算法来提高语音翻译任务的性能；STEMM (Fang et al., 2022)通过混合语音和转录文本序列，并应用正则化方法来学习更好的语音表示；ConST (Ye et al., 2022)利用对比学习来缩小模态差异；FCGCL (Zhang et al., 2022a)利用多粒度对比学习来缩小模态差异；SpeechUT (Zhang et al., 2022b)利用有监督和无监督数据进行联合预训练，并通过统一的编码器来处理语音和文本模态。

[3]https://dl.fbaipublicfiles.com/hubert/hubert_base_ls960.pt
[4]https://github.com/pytorch/fairseq
[5]https://github.com/mjpost/sacrebleu, signature: nrefs:1 | bs:1000 | seed:12345 | case:mixed | eff:no | tok:13a | smooth:exp | version:2.0.0


(c) 2024 中国中文信息学会计算语言学专业委员会

| 模型 | BLEU | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | EN-DE | EN-FR | EN-RU | EN-ES | EN-IT | EN-RO | EN-PT | EN-NL | Avg. |
| 基于Wav2Vec2.0 | | | | | | | | | |
| Chimera (Han et al., 2021) | 27.1 | 35.6 | 17.4 | 30.6 | 25.0 | 24.0 | 30.2 | 29.2 | 27.4 |
| XSTNet (Ye et al., 2021) | 27.1 | 38.0 | 18.5 | 30.8 | 26.4 | 25.7 | 32.4 | 31.2 | 28.8 |
| STEMM (Fang et al., 2022) | 28.7 | 37.4 | 17.8 | 31.0 | 25.8 | 24.5 | 31.7 | 30.5 | 28.4 |
| ConST (Ye et al., 2022) | 28.3 | 38.3 | 18.9 | 32.0 | 27.2 | 25.6 | 33.1 | 31.7 | 29.4 |
| FCGCL (Zhang et al., 2022a) | 28.7 | 37.5 | 19.1 | 31.2 | 26.5 | 26.0 | 32.1 | 31.0 | 29.0 |
| SpeechUT (Zhang et al., 2022b) | **30.1** | **41.4** | - | **33.6** | - | - | - | - | - |
| W2V2-ST | 27.1 | 37.7 | 18.5 | 31.8 | 26.2 | 24.4 | 30.7 | 30.4 | 28.4 |
| 本文方法 | 28.8* | 39.2* | **19.9*** | 33.0* | **27.3*** | **26.4*** | **33.2*** | **32.2*** | **30.0** |
| 基于HuBERT | | | | | | | | | |
| CMOT (Zhou et al., 2023) | 29.0 | 39.5 | 19.2 | 32.8 | 27.5 | 26.0 | 33.5 | 32.1 | 30.0 |
| CRESS (Fang and Feng, 2023) | 29.4 | 40.1 | 19.7 | 33.2 | 27.6 | **26.4** | 33.6 | 32.3 | 30.3 |
| DUB (Zhang et al., 2023a) | 26.2 | 35.3 | - | 30.4 | - | - | - | - | - |
| IMTL-KD (Zhang et al., 2023c) | **29.7** | **41.1** | - | **33.9** | - | - | - | - | - |
| HuBERT-ST | 27.5 | 38.1 | 19.0 | 32.3 | 26.5 | 25.1 | 31.8 | 30.8 | 28.9 |
| 本文方法 | 29.2* | 40.0* | **20.1*** | 33.3* | **27.7*** | **26.4*** | **33.8*** | **32.4*** | **30.4** |

表 2. 在MuST-C测试集tst-COMMON上与其它方法性能比较，*表示较基线模型(即W2V2-ST或HuBERT-ST)的性能提升在BLEU值方面具有统计学显著性($p < 0.01$) (Koehn, 2004)

| # | 模型 | BLEU |
|---|---|---|
| 1 | ST | 25.0 |
| 2 | CRESS | **26.7** |
| 3 | 本文方法 | 26.6* |

表 3. 在CoVoST 2英语-德语测试集上性能对比，*表示较#1的性能提升在BLEU值方面具有统计学显著性($p < 0.01$) (Koehn, 2004)

- 基于HuBERT的模型：CMOT (Zhou et al., 2023)在STEMM的基础上，利用最优传输理论自动找到语音和转录文本之间的对齐，并利用对齐信息来混合语音和转录文本序列；CRESS (Fang and Feng, 2023)把模态差异和机器翻译任务的暴露偏差问题关联起来，提出目标端计划采样、正则化方法、令牌级自适应训练来缓解模态差异。DUB (Zhang et al., 2023a)使用离散单元代替连续的语音序列，并通过反向翻译技术提高语音翻译任务的性能；IMTL-KD (Zhang et al., 2023c)分析了多任务训练方法中不同任务之间的一致性，为语音翻译任务提出了改进的多任务训练方法，该方法通过减少语音和文本的长度和表征差异来缩小模态差异。

参考Fang等人 (2022)的做法，本文使用相同的模型结构实现了类似的基线模型HuBERT-ST，HuBERT-ST仅仅对语音翻译任务进行模型参数优化。为了同基于Wav2vec2.0的模型进行比较，本文基于Wav2vec2.0[6]实现了本文的方法和基线系统W2V2-ST。实验结果如表2所示，从表2我们可以看出，

- 无论是基于Wav2Vec2.0还是基于HuBERT，本文模型明显优于相应的基线模型。相比于基线模型，本文模型在八个语向的BLEU平均分别提升1.6个点和1.5个点。
- 相比于其它端到端语音翻译模型，本文模型平均而言取得了最好的性能[7]。

## 5 分析与讨论

### 5.1 CoVoST 2英语-德语方向性能

参照CRESS (Fang and Feng, 2023)做法，本文除了在MuST-C数据集的八个语向上进行实验之外，还在更大的数据集CoVoST 2 (Wang et al., 2020a)上测试本文方法的性能。CoVoST

[6]https://dl.fbaipublicfiles.com/fairseq/wav2vec/wav2vec_small.pt
[7]由于SpeechUT和IMTL-KD大部分语向性能并未给出，此处不考虑SpeechUT和IMTL-KD


(c) 2024 中国中文信息学会计算语言学专业委员会

| # | 模型 | BLEU |
|---|---|---|
| 1 | 本文方法 | **29.2*** |
| 2 | −GACMLM | 28.9* |
| 3 | −SATT | 28.5* |
| 4 | −GACMLM−SATT | 27.5 |

表 4. 在MuST-C英语-德语测试集tst-COMMON上的消融实验结果，*表示较#4的性能提升在BLEU值方面具有统计学显著性($p < 0.01$) (Koehn, 2004)

| # | 模型 | BLEU |
|---|---|---|
| 1 | ST | 27.5 |
| 2 | ST+TT | 28.6* |
| 3 | ST+SATT | **28.9*** |

表 5. 在MuST-C英语-德语测试集tst-COMMON上性能对比，*表示较#1的性能提升在BLEU值方面具有统计学显著性($p < 0.01$) (Koehn, 2004)

2是一个大型多语种语音翻译数据集，涵盖从21种语音翻译成英语和从英语语音翻译成15种语言的内容。本文在英语-德语方向进行实验，该方向包含430小时的语音，并带有对应的转录文本和译文。本文使用开发集进行验证，使用测试集进行评估，使用与MuST-C相同的预处理、模型配置和超参数(见4.1)。结果如表3所示，表中"ST"表示基线模型，其仅在语音翻译任务上训练。从表中可以看出，本文方法的性能明显优于基线系统(+1.6)并且与CRESS相比具有可竞争的性能，表明本文的方法在不同数据集上的有效性。

### 5.2 消融分析

为了探究不同部分对总体模型的贡献，本文在英语-德语方向测试集tst-COMMON上进行消融实验。实验结果如表4所示。从表中可以看出，在移除全局感知条件掩码语言建模(GACMLM)任务后，BLEU下降了0.3个点，表明文本的全局上下文信息有利于语音翻译任务。在移除语音增强的文本翻译(SATT)任务后，BLEU下降了0.7个点，表明该任务对于总体模型影响较大。在移除本文提出的两个辅助任务后，BLEU急剧下降至27.5(-1.7)，表明在多任务学习框架下，联合使用本文提出的两个新颖辅助任务能够有效提升语音翻译任务的性能。

### 5.3 文本翻译中语音的作用

为了探究语音增强的文本翻译任务中语音起到的效果。本文在MuST-C测试集tst-COMMON上做了进一步的实验，实验结果如表5所示，表中各个模型均已经过第一阶段外部机器翻译语料预训练，具体代表的含义如下所示:

- ST：表示模型仅在语音翻译任务上训练(即HuBERT-ST)。
- ST+TT：表示模型在语音翻译任务、文本翻译任务和模态正则化损失上进行联合训练。
- ST+SATT：表示模型在语音翻译任务和语音增强的文本翻译任务上进行联合训练。

比较表中#1和#2，我们可以看出，在移除文本翻译任务后，BLEU下降了1.1，表明在多任务学习框架下文本翻译任务的加入能够有效提升语音翻译任务的性能。比较表中#2和#3，我们可以看出，在源端对转录文本进行自注意力计算时融入语音模态信息有利于语音翻译任务的性能提升(+0.3)，验证了本文提出的语音增强的文本翻译任务的有效性。

### 5.4 知识蒸馏权重$\alpha$和掩码概率$p$对语音翻译性能的影响

本文提出全局感知条件掩码语言建模任务来改善语音翻译任务的训练。为了探究公式(15)中知识蒸馏权重$\alpha$对语音翻译性能的影响，本文在MuST-C英语-德语验证集上进行网格搜索，令$\alpha \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$，实验结果如图2(左)所示。从图中可以看出，当$\alpha = 0.5$时，MuST-C验证集上性能最佳。为了探究条件掩码语言模型中单词掩


(c) 2024 中国中文信息学会计算语言学专业委员会

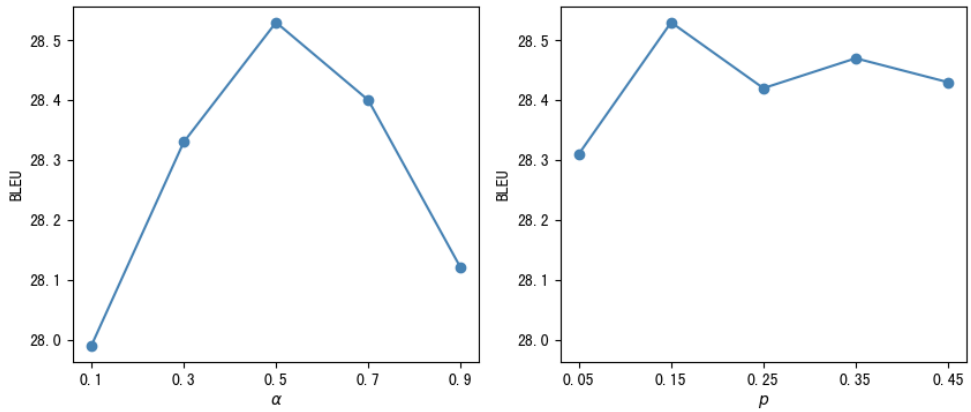



图 2. 在MuST-C英语-德语验证集上知识蒸馏权重$\alpha$(左)、掩码概率$p$(右)对语音翻译性能的影响

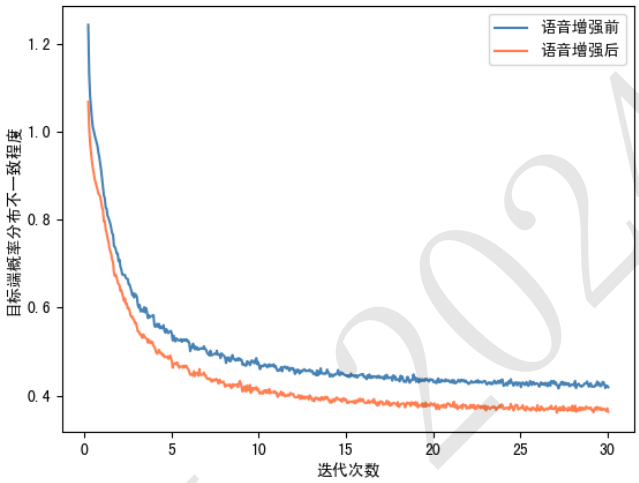



图 3. 在MuST-C英语-德语训练过程中目标端预测概率分布的不一致程度

码概率$p$对最终语音翻译性能的影响，本文同样在MuST-C英语-德语验证集上进行网格搜索，令$p \in \{0.05, 0.15, 0.25, 0.35, 0.45\}$，实验结果如图2(右)所示。从图中可以看出，当$p = 0.15$时，MuST-C验证集上性能最佳。因此，本文其它实验均设置$\alpha = 0.5$，$p = 0.15$。

### 5.5 模态差异

本文利用公式(9)来衡量语音翻译任务和语音增强的文本翻译任务预测的目标端概率分布的不一致程度。为了探究语音增强前后目标端概率分布不一致程度的变化状况，本文在英语-德语方向的训练过程中绘制了语音增强前后的不一致程度变化曲线，如图3所示。其中，“语音增强前”表示模型在语音翻译任务、文本翻译任务和模态正则化损失上训练得到(即ST+TT)，“语音增强后”表示模型在语音翻译任务和本文提出的语音增强的文本翻译任务上训练得到(即ST+SATT)。从图中可以看出，在相同迭代次数下，进行语音增强后，语音翻译任务和文本翻译任务对应的目标端概率分布不一致性程度更低，进一步表明了语音增强的有效性。

为了判断本文提出的辅助任务是否缓解了语音和文本之间的模态差异，本文使用余弦相似度来衡量二者的模态差异。具体而言，针对语音翻译任务和语音增强的文本翻译任务，我们分别获取其经过翻译编码器后的输出$\mathbf{e_{mix}}$和$\mathbf{a}$，接着在时间维度进行平均池化来获取相应的语音表示和文本表示并计算二者的余弦相似度。本文在MuST-C英语-德语测试集tst-COMMON进行实验。实验结果如表6所示，表中各个模型代表的含义与5.3节相同。从表中#1和#2可以看出，对文本翻译任务进行语音增强后能够将语音表示和文本表示的余弦相似度提升2.45%。从


(c) 2024 中国中文信息学会计算语言学专业委员会

| # | 模型 | 余弦相似度（%） |
|---|---|---|
| 1 | ST+TT | 86.20 |
| 2 | ST+SATT | 88.65 |
| 3 | 本文方法 | **90.32** |

表 6. 在MuST-C英语-德语测试集tst-COMMON上模态差异对比

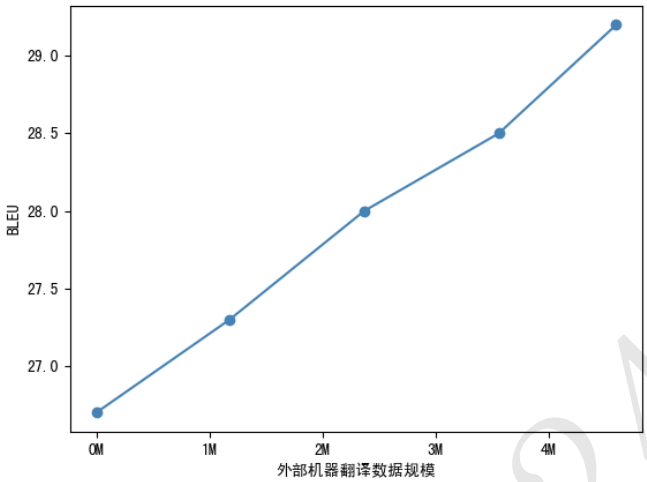



图 4. 在MuST-C英语-德语测试集tst-COMMON上外部机器翻译数据规模与性能对比

表中#2和#3可以看出，本文提出的全局感知条件掩码语言建模任务能够把余弦相似度再次提升1.67%。余弦相似度的提升验证了本文提出的辅助任务能够缓解语音和文本的模态间差异。

| # | 模型 | BLEU |
|---|---|---|
| 1 | TT | **34.0** |
| 2 | ST | 27.2 |
| 3 | ST+SATT | 33.5 |
| 4 | 本文方法 | **34.0** |

表 7. 在MuST-C英语-德语测试集tst-COMMON上的文本翻译任务性能对比

### 5.6 文本翻译性能

在多任务学习框架下，语音翻译任务性能的提升可能会导致文本翻译任务性能出现下降。为了探究本文提出的辅助任务是否会损害文本翻译任务的性能，本文在MuST-C英语-德语测试集tst-COMMON上进行实验。实验结果如表7所示，表中“TT”表示模型仅在文本翻译任务上训练，其它模型代表的含义与5.3节相同。从表中#1和#2可以看出，模型仅在语音翻译任务上训练后，文本翻译任务的性能出现大幅下降(-6.8)。从表中#1和#3可以看出，在应用本文提出的语音增强的文本翻译任务后，文本翻译任务的性能仅出现略微下降(-0.5)。最终，再加入全局感知条件掩码语言建模任务后，原本文本翻译任务的性能依然得到了保留。

### 5.7 外部机器翻译数据对语音翻译性能的影响

为了探究外部机器翻译数据对本文方法的影响，本文在MuST-C数据集的英语-德语方向进行了实验。将英语-德语方向的460M外部机器翻译数据分成五等份，并逐步增加外部机器翻译数据的使用量。针对每个增量量级分别在tst-COMMON测试集上进行评估。实验结果如图4所示。从图中可以看出，随着外部机器翻译数据使用量的增加，本文方法的BLEU值持续提高。


(c) 2024 中国中文信息学会计算语言学专业委员会

## 6 总结

本文在多任务学习框架下提出两个新颖辅助任务来改善语音翻译任务的训练。其中，语音增强的文本翻译任务通过使转录文本关注语音模态的信息，缓解了语音和文本之间的模态差异，从而提升了语音翻译任务的性能。全局感知的条件掩码语言建模任务使模型能够同时建模转录文本和译文，学习文本自身的全局上下文信息并利用这些信息指导翻译模型的训练。相比于其它最新的端到端语音翻译模型，本文的方法能够在MuST-C数据集上取得有竞争力的性能。相关的分析实验表明本文方法的有效性，它能够缓解语音和文本之间的模态差异，同时不损害文本翻译任务的性能。

## 参考文献

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proceedings of NIPS*, volume 33, pages 12449–12460.

Parnia Bahar, Albert Zeyer, Ralf Schlüter, and Hermann Ney. 2019. On using specaugment for end-to-end speech translation. In *Proceedings of IWSLT*.

Yen-Chun Chen, Zhe Gan, Yu Cheng, Jingzhou Liu, and Jingjing Liu. 2020. Distilling knowledge learned in BERT for text generation. In *Proceedings of ACL*, pages 7893–7905, July.

Nanxin Chen, Shinji Watanabe, Jesús Villalba, Piotr Żelasko, and Najim Dehak. 2021. Non-autoregressive transformer for speech recognition. *IEEE Signal Processing Letters*, 28:121–125.

Xuxin Cheng, Qianqian Dong, Fengpeng Yue, Tom Ko, Mingxuan Wang, and Yuexian Zou. 2023. M3st: Mix at three levels for speech translation. In *Proceedings of ICASSP*, pages 1–5.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a Multilingual Speech Translation Corpus. In *Proceedings of NAACL*, pages 2012–2017.

Qianqian Dong, Feng Wang, Zhen Yang, Wei Chen, Shuang Xu, and Bo Xu. 2019. Adapting translation models for transcript disfluency detection. In *Proceedings of AAAI*, pages 6351–6358.

Qingkai Fang and Yang Feng. 2023. Understanding and bridging the modality gap for speech translation. In *Proceedings of ACL*, pages 15864–15881.

Qingkai Fang, Rong Ye, Lei Li, Yang Feng, and Mingxuan Wang. 2022. STEMM: Self-learning with speech-text manifold mixup for speech translation. In *Proceedings of ACL*, pages 7050–7062.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. 2019. Mask-predict: Parallel decoding of conditional masked language models. In *Proceedings of EMNLP-IJCNLP*, pages 6112–6121.

Chi Han, Mingxuan Wang, Heng Ji, and Lei Li. 2021. Learning shared semantic space for speech-to-text translation. In *Findings of ACL-IJCNLP*, pages 2214–2225.

Yuchen Han, Chen Xu, Tong Xiao, and Jingbo Zhu. 2023. Modality adaption or regularization? a case study on end-to-end speech translation. In *Proceedings of ACL*, pages 1340–1348.

Yosuke Higuchi, Shinji Watanabe, Nanxin Chen, Tetsuji Ogawa, and Tetsunori Kobayashi. 2020. Mask CTC: Non-Autoregressive End-to-End ASR with CTC and Mask Predict. In *Proceedings of INTERSPEECH*, pages 3655–3659.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE ACM Trans. Audio Speech Lang. Process.*, 29:3451–3460.

Wuwei Huang, Dexin Wang, and Deyi Xiong. 2021. AdaST: Dynamically adapting encoder states in the decoder for end-to-end speech-to-text translation. In *Findings of ACL-IJCNLP*, pages 2539–2545.


(c) 2024 中国中文信息学会计算语言学专业委员会

Hirofumi Inaguma, Yosuke Higuchi, Kevin Duh, Tatsuya Kawahara, and Shinji Watanabe. 2021. Orthros: non-autoregressive end-to-end speech translation with dual-decoder. In *Proceedings of ICASSP*, pages 7503–7507.

Ye Jia, Melvin Johnson, Wolfgang Macherey, Ron J. Weiss, Yuan Cao, Chung-Cheng Chiu, Naveen Ari, Stella Laurenzo, and Yonghui Wu. 2019. Leveraging weakly supervised data to improve end-to-end speech-to-text translation. In *Proceedings of ICASSP*, pages 7180–7184.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of ICLR*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of EMNLP*, pages 388–395.

Tsz Kin Lam, Shigehiko Schamoni, and Stefan Riezler. 2021. Cascaded models with cyclic feedback for direct speech translation. In *Proceedings of ICASSP*, pages 7508–7512.

Tsz Kin Lam, Shigehiko Schamoni, and Stefan Riezler. 2022. Sample, translate, recombine: Leveraging audio alignments for data augmentation in end-to-end speech translation. In *Proceedings of ACL*, pages 245–254.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL*, pages 48–53.

Siqi Ouyang, Rong Ye, and Lei Li. 2023. WACO: Word-aligned contrastive learning for speech translation. In *Proceedings of ACL*, pages 3891–3907.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318.

Juan Pino, Liezl Puzon, Jiatao Gu, Xutai Ma, Arya D. McCarthy, and Deepak Gopinath. 2019. Harnessing indirect training data for end-to-end automatic speech translation: Tricks of the trade. In *Proceedings of IWSLT*.

Juan Miguel Pino, Qiantong Xu, Xutai Ma, Mohammad Javad Dousti, and Yun Tang. 2020. Self-training for end-to-end speech translation. In *Proceedings of INTERSPEECH*, pages 1476–1480.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

Yi Ren, Jinglin Liu, Xu Tan, Chen Zhang, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2020. SimulSpeech: End-to-end simultaneous speech to text translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3787–3796.

Matthias Sperber and Matthias Paulik. 2020. Speech translation and the end-to-end promise: Taking stock of where we are. In *Proceedings of ACL*, pages 7409–7421.

Yun Tang, Juan Pino, Xian Li, Changhan Wang, and Dmitriy Genzel. 2021. Improving speech translation by understanding and learning from the auxiliary text translation task. In *Proceedings of ACL-IJCNLP*, pages 4252–4261.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of NIPS*, pages 5998–6008.

Changhan Wang, Anne Wu, and Juan Miguel Pino. 2020a. Covost 2: A massively multilingual speech-to-text translation corpus. *CoRR*, abs/2007.10310.

Chengyi Wang, Yu Wu, Shujie Liu, Zhenglu Yang, and Ming Zhou. 2020b. Bridging the gap between pre-training and fine-tuning for end-to-end speech translation. In *Proceedings of AAAI*, pages 9161–9168.

Changhan Wang, Anne Wu, Juan Pino, Alexei Baevski, Michael Auli, and Alexis Conneau. 2021. Large-scale self-and semi-supervised learning for speech translation. In *Proceedings of INTERSPEECH*, pages 2242–2246.

CCL 2024

Chen Xu, Rong Ye, Qianqian Dong, Chengqi Zhao, Tom Ko, Mingxuan Wang, Tong Xiao, and Jingbo Zhu. 2023. Recent advances in direct speech-to-text translation. In *Proceedings of IJCAI*, pages 6796–6804.

Rong Ye, Mingxuan Wang, and Lei Li. 2021. End-to-End Speech Translation via Cross-Modal Progressive Training. In *Proceedings of INTERSPEECH*, pages 2267–2271.

Rong Ye, Mingxuan Wang, and Lei Li. 2022. Cross-modal contrastive learning for speech translation. In *Proceedings of NAACL*, pages 5099–5113.

Wenbiao Yin, Zhicheng Liu, Chengqi Zhao, Tao Wang, Jian Tong, and Rong Ye. 2023. Improving speech translation by fusing speech and text. In *Findings of EMNLP*, pages 6262–6273.

Pei Zhang, Niyu Ge, Boxing Chen, and Kai Fan. 2019. Lattice transformer for speech translation. In *Proceedings of ACL*, pages 6475–6484.

Biao Zhang, Ivan Titov, Barry Haddow, and Rico Sennrich. 2020a. Adaptive feature selection for end-to-end speech translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2533–2544.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. 2020b. Improving massively multilingual neural machine translation and zero-shot translation. In *Proceedings of ACL*, pages 1628–1639.

Hao Zhang, Nianwen Si, Yaqi Chen, Zhen Li, Tong Niu, Xukui Yang, and Dan Qu. 2022a. FCGCL: Fine- and coarse-granularity contrastive learning for speech translation. In *Findings of EMNLP*, pages 3048–3059.

Ziqiang Zhang, Long Zhou, Junyi Ao, Shujie Liu, Lirong Dai, Jinyu Li, and Furu Wei. 2022b. SpeechUT: Bridging speech and text with hidden-unit for encoder-decoder based speech-text pre-training. In *Proceedings of EMNLP*, pages 1663–1676.

Dong Zhang, Rong Ye, Tom Ko, Mingxuan Wang, and Yaqian Zhou. 2023a. DUB: Discrete unit back-translation for speech translation. In *Findings of ACL*, pages 7147–7164.

Linlin Zhang, Kai Fan, Boxing Chen, and Luo Si. 2023b. A simple concatenation can effectively improve speech translation. In *Proceedings of ACL*, pages 1793–1802.

Yuhao Zhang, Chen Xu, Bei Li, Hao Chen, Tong Xiao, Chunliang Zhang, and Jingbo Zhu. 2023c. Rethinking and improving multi-task learning for end-to-end speech translation. In *Proceedings of EMNLP*, pages 10753–10765.

Chulun Zhou, Fandong Meng, Jie Zhou, Min Zhang, Hongji Wang, and Jinsong Su. 2022. Confidence based bidirectional global context aware training framework for neural machine translation. In *Proceedings of ACL*, pages 2878–2889.

Yan Zhou, Qingkai Fang, and Yang Feng. 2023. CMOT: Cross-modal mixup via optimal transport for speech translation. In *Proceedings of ACL*, pages 7873–7887.

CCL 2024

# 基于隐性句逗号识别的汉语长句机器翻译

张文娟　　李熳佳　　冯文贺*
广东外语外贸大学外国语言学及应用语言学研究中心/语言工程与计算实验室
{20211050028, 20221050034, wenhefeng@}@gdufs.edu.cn

## 摘要

长句翻译一直是机器翻译的难题。本文根据汉语中相当数量的逗号（句内标点）和句号（句间标点）可相互转化的特点，提出“隐性句号”（可转化为句号的逗号）和“隐性逗号”（可转化为逗号的句号）概念，并实现其自动识别，以将汉语长句变为短句用于汉英机器翻译。为此，首先通过人工与半监督学习结合方法构建了一个隐性句逗数据集，实现了基于预训练模型的隐性句逗识别方法，其中性能最好的Hierarchical BERT作为后续应用模型。进而，实现了基于隐性句逗识别的汉英机器翻译方法。在WMT2018（新闻）和WMT2023（文学）测试语料上基于预训练机器翻译模型的实验表明，对于汉语长句的英译，本文方法相比基准翻译的BLEU值整体有所提高，而且在相对稳健机器翻译模型上，呈现为句子越长本文方法效果越明显。

**关键词：** 机器翻译 ； 长句翻译 ； 隐性句逗号

# Machine translation of Chinese long sentences based on recognition of implicit period and comma

**Wenjuan Zhang　　Manjia Li　　Wenhe Feng**
Lab of language engineering and computing
Center for Linguistics and Applied Linguistics
Guangdong University of Foreign Studies
{20211050028, 20221050034, wenhefeng@}@gdufs.edu.cn

## Abstract

The translation of long sentences has always been a difficult problem for machine translation. In this paper, based on the feature that a considerable number of commas (intra-sentence punctuation) and periods (inter-sentence punctuation) in Chinese text can be transformed into each other, we propose the concepts of "implicit period" (comma that can be transformed into period) and "implicit comma" (period that can be transformed into comma), and realize their automatic recognition to transform Chinese long sentences into short sentences for Chinese-English machine translation. In this paper, a dataset of implicit period and comma is constructed by combining manual and semi-supervised learning methods, and an implicit period and comma recognition method is realized based on a pre-trained model, in which Hierarchical BERT, which has the best performance, is used as the subsequent application model. In this paper,

*通讯作者(冯文贺）：wenhefeng@gdufs.edu.cn
©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版
**项目基金**:广东省社科基金项目（GD24CWY11）；广东省教育厅GK特色创新项目（2023WTSCXO17）；广东外语外贸大学外国语言学及应用语言学研究中心语言与人工智能实验室招标课题（LAI202304）

a Chinese-English machine translation method based on implicit period and comma recognition is realized. The experiments based on pre-trained machine translation models on the WMT2018 (News) and WMT2023 (Literature) corpus show that for the English translation of long Chinese sentences, the method in this paper improves the BLEU value compared to the benchmark translation as a whole, and the effect of the method in this paper is more obvious the longer the sentence is for the relatively robust machine translation model.

**Keywords:** machine translation , long sentence translation , implicit period and comma

## 1 引言

篇章翻译是当前制约机器翻译技术性能的一个突出问题，其困难集中体现于长句翻译上 (Koehn and Knowles, 2017)。原因在于长句一般由多个小句（clause)构成，而不同语言间小句及其间结构有重要差异。如表现在汉英语言间，汉语小句无主从之别，长句多流水句，小句间“可断可连” (吕叔湘, 1979)，反映在标点上即逗号（句内标点）、句号（句间标点）可相互转换，句子边界相对模糊；而英语小句主从分明，相互转化性弱，句逗分明，句子边界相对清晰。小句及其间结构与句子边界差异，自然会引发双语长句的翻译问题。长句机器翻译困难也有其计算机制原因。如对于基于循环神经网络的翻译模型，维护长距离依赖关系时可能难以记住大量上下文信息；对于基于注意力机制的翻译模型，会造成注意力分散到更多信息点，导致难以持续聚焦于最关键信息。

对于长句翻译问题，一种解决的思路就是将其化为短句翻译。问题是如何合理将长句化为短句。本文考虑，可从汉语句逗号中有大量可相互转化而不影响原义表达的特点入手，解决汉语长句的翻译问题。例如：

**例1** ①少年姓孙，②属马[，]③比小水小着一岁，④个头也没小水高，⑤人却本分实诚。（贾平凹《浮躁》）

E1：①This boy,a member of Sun Family, ②was born in the year of the horse[.]③Although he was a year younger ④and a head shorter than water girl, ⑤he was honest and sincere.（Goldblatt 1991）

E2：①The young Sun, ②a horse, ③is one year younger than Xiaoshui, ④and the individual is no higher than Xiaoshui, ⑤but the person is sincere.（有道翻译2023.11.18）

E3：①The young man’s surname is Sun, ②he was born in the year of the horse, ③and is one year younger than Xiao Shui. ④He is not as tall as Xiao Shui, ⑤but he is honest and earnest.（ChatGPT4 2023.12.10）

（说明：数字序号代表汉英小句序号；汉语隐性句逗号与其对应英译标点用红色[]标出，下同。原文例句、E1及其小句切分采用自(冯文贺, 2019)）

例1汉语复句包含5个小句。语义主题上，小句①②描述少年的个人特征；小句③④和小句⑤虽然也描述少年的个人特征，但却有了对比人物“小水”。据此，前两个小句和后三个小句所表达语义主题是有差异的，其间语义关系也远近有别。也因此，可将小句②后逗号改为句号，而且原文小句间的语义关系远近及具体逻辑语义等不变。本文将此类可转化为句号而不改变原文小句间语义关系远近及逻辑语义的逗号视为”隐性句号”。同理，汉语文本中也存在句号可变为逗号的情况，本文将此类句号视为“隐性逗号”。

将汉语长句变为短句，相比可以实现更好翻译。如专业译者（E1）在小句②后断句翻译，更好翻译表达了源语小句间的语义远近关系，句子的内部结构也更符合英文习惯。相比之下，机器译文E2中句逗号与源语一致，小句间的语义关系不明，句子结构也显得冗长，不符合英文结构习惯。但是，如果不能恰当化长句为短句，也可能导致不良翻译。如机器译文E3中在小句③后断句翻译，根本上改变源语小句间的结构关系，其中割裂了小句③④间密切的并列关系（由连接词“也”表示），隔断了小句⑤对小句③④整体的转折关系（由此也导致译文连接


(c) 2024 中国中文信息学会计算语言学专业委员会

词“but”不能准确翻译原文连接词“却”的管辖范围）。可以说，长句化短后翻译，在根本上会影响句子内外的结构组织，影响是全局性的。

本文考虑，如果能够识别汉语文本中隐性句逗号，就可能将相当一部分长句经切分重组为较短句子，而经此处理后再进行机器翻译，就可能得到更好的翻译效果。基于此，本文提出了基于隐性句逗号识别的汉语长句机器翻译方法。为此，本文首先构建了汉语隐性句逗号数据集，并实现了其识别模型；进而将隐性句逗号识别模型用于机器翻译的汉语源语数据预处理，以使机器翻译获得更好长句翻译效果。在WMT2018(新闻）和WMT2023（文学）翻译测试集上的实验结果表明，本文方法可以显著提高汉语长句的机器翻译性能，而且长句越长，效果越明显。

**本文主要贡献：**（1）提出并实现汉语的隐性句逗号识别：针对汉语中句逗号可相互转化现象，提出“隐性句逗号”概念，构建相应数据集，并实现其自动识别技术，为长句化短提供方法。（2）提出并实现基于隐性句逗识别的长句机器翻译方法：结合隐性句逗号识别模型与机器翻译模型，在翻译前将源语汉语长句化短，显著提高了汉语长句的机器翻译质量。

## 2 相关工作

现有机器翻译模型处理短句表现良好，但对于长句，往往无法给出优质翻译。为改善长句翻译效果，主要进行了两类研究：一类是进行篇章级机器翻译建模，综合解决包括长句翻译在内的篇章翻译问题；另一类是专门针对长句的机器翻译研究。

在篇章机器翻译建模研究中，一般既关注句内词语间结构关系，也关注上下文句子间的一致性、连贯性、结构层次、衔接性等信息(Tan et al., 2019; Chen et al., 2020; Guo et al., 2022; 贾爱鑫et al., 2024)。由于关注更多上下文，一定程度上有利于缓解长句翻译问题。在篇章机器翻译建模中，有研究特别注意到句长的影响。如有研究指出篇章级机器翻译中源语和目标语的句长偏差会导致翻译质量下降；提出动态采样训练数据，以确保不同序列长度可均匀分布；引入长度归一化注意力机制，以使模型聚焦于目标信息，缓解处理较长序列时的注意力偏离问题；提出在解码过程中采用滑动窗口策略，以在不超过最大序列长度的前提下整合更多上下文信息(Zhang et al., 2023)。然而，根本上篇章机器翻译建模并不着重于解决由于（汉英）双语句子边界差异带来的长句翻译问题。

在专门针对长句的机器翻译研究中，一般将长句化为较短的语言单位再翻译。在传统机器翻译（基于规则、基于统计）下，不同语言的翻译上均有研究试图将长句划分较短的语言单位再翻译，一般是利用一些语言特征，如句法模板、结构层次、小句、连接词、标点等，将长句划分为更短的短语、小句、结构片段等后，先翻译再组合(Oliveira et al., 2010; Goh and Sumita, 2011; Yin et al., 2012; Hung et al., 2012; Pouget-Abadie et al., 2014)。在汉英翻译中，也有一些研究尝试利用标点符号、关系代词、层次结构等长句划分为简单句、子句等后再进行翻译(黄河燕and 陈肇雄, 2002; 李幸and 宗成庆, 2006; Xiong et al., 2009)。在神经机器翻译技术框架下，也有研究考虑将长句化短后分别翻译再组合。如有研究在汉英翻译中引入一个拆分和重新排序模型来共同检测源语长句的最佳分割点序列；进而，将每个源语子句由NMT系统独立转换为目标子句，并将翻译的目标子句连接起来形成长句的最终翻译(Kuang and Xiong, 2016)。有研究提出一种提取双语短语的方法来构建短语对齐的双语语料库，并实现了一种长句预处理技术，以此切分长句为短语，以解决长句翻译(Tien and Minh, 2019)。长句分割与句子边界相关，有研究发现发现句子边界分割对口语翻译质量影响显著，提出一种数据增强策略，即在训练过程中将模型暴露于各种边界分割错误中，以提高神经机器翻译系统对句子边界分割错误处理的鲁棒性和机器翻译的准确性(Li et al., 2021)。然而这些研究只是一般地将长句化为较短的语言单位再翻译，而不考虑长句化短后是否改变了源语长句内外，小句间语义关系的远近与逻辑类别等。但事实上，随意切断长句后翻译可能导致原文小句间的逻辑语义结构的改变（如例1的E3）。为此，本文基于汉语部分句逗号可相互转化的特点，提出“隐性句逗号”概念，实现其机器识别，并用以解决汉语长句的机器翻译难题，并特别关注机器翻译中长句内外小句间语义关系的远近与逻辑类别等是否得到了准确翻译。

## 3 隐性句逗号识别

本文构建了汉语隐性句逗号数据集，并实现了基于预训练语言模型的隐性句逗号识别方法。考虑到原始文本中句逗号的数据不平衡（句号少，逗号多），及隐性句逗号的不平衡，和


(c) 2024 中国中文信息学会计算语言学专业委员会

| 类型 | 汉语文本 | 英语文本 |
|---|---|---|
| 隐性句号 | **S1**祥子，**S2**在与“骆驼”这个外号发生关系以前，**S3**是个较比有自由的洋车夫[，]**S4**这就是说，**S5**他是属于年轻力壮，**S6**而且自己有车的那一类。 | he gained the nickname Camel), he was a relatively independent rickshaw man[.]That is to say, he belonged to the young, vigorous set and owned his own rickshaw. |
| 隐性逗号 | **S1**从风里雨里的咬牙，**S2**从饭里茶里的自苦，**S3**才赚出那辆车[。]**S4**那辆车是他的一切挣扎与困苦的总结果与报酬，**S5**象身经百战的武士的一颗徽章。 | By gritting his teeth through wind and rain and scrimping on food and tea, he finally put enough aside to buy it[,] a tangible reward for his struggles and his suffering, like a medal for valor. |

表 1: 人工标注隐性句逗号数据集示例

| | 数据集划分 | 数据量 | 平均长度 | 总词数 |
|---|---|---|---|---|
| 隐性句号 | train | 15,526 | 118 | 1,365,404 |
| | valid | 1,940 | 118 | 170,785 |
| | test | 1,941 | 115 | 166,151 |
| 隐性逗号 | train | 11,552 | 28 | 438,107 |
| | valid | 1,444 | 27 | 53,431 |
| | test | 1,444 | 29 | 55,328 |

表 2: 半监督扩充数据集统计

| | 文本长（逗号数） | 样本数 | 字数 | 平均长度 | 词表大小 |
|---|---|---|---|---|---|
| WMT2018 新闻 | $< 3$ | 34021 | 1599099 | 47 | 3783 |
| | $= 3$ | 3643 | 234,953 | 65 | 2,834 |
| | $= 4$ | 1091 | 82,182 | 75 | 2,332 |
| | $= 5$ | 265 | 22,961 | 87 | 1,692 |
| | $> 5$ | 100 | 10,198 | 102 | 1,368 |
| WMT2023文学 | $> 3$ | 552 | 34,211 | 62 | 1,754 |

表 3: 机器翻译数据统计

预训练语言模型中本身句逗号知识的不平衡，本文专门构建了一个只包含隐性句逗号的数据集。基于该数据集上，我们训练实现了最优隐性句逗号识别模型，该模型可以相对集中的反映隐性句逗号的特征差异。在机器翻译中，该模型将用于预处理源语汉语文本，由于该模型并未关注真句逗号，其识别结果将与现实文本中句逗号的进行一致性对比调正后作为源语文本预处理结果（见4.1节），输入机器翻译模型进行翻译。

**隐性句逗号数据集**：为了训练与验证隐性句逗号模型，本文构建了隐性句逗号数据集。首先，在不同体裁的汉语文本上标注共计3000条隐性句逗号；然后通过self-training半监督学习方法大规模扩充数据集。

人工标注由汉语母语者实施，通过两种方式实现。**标注方式1**：标注者根据母语者的语感直接对汉语文本标注。基本判断标准：句逗号相互转变后，语法合理、且不改变原句所含逻辑语义关系的，为隐性句逗号。**标注方式2**：参照汉英翻译标注。标准为：在经典汉英翻译平行语料上，如果英译文本为句号断句，而汉语文本的对应标点处为逗号，则认定该汉语标点为隐性句号；隐性逗号的确定方法同理。具体做法如表 1所示，其中红色标注出了隐性句逗号，连同其左右各一个标点句（用S1、S2等标注）(宋柔, 2022)构成一条数据。如其中的隐性句号数据由S3-S4构成，隐性逗号数据由S3-S4构成。

人工标注共计3000条隐性句逗号样本，其中包含1847条隐性句号标注，1253条隐性逗号标注。随后，采用了self-training半监督学习方法大规模扩充数据集。先将这些标注样本作为初始数据，然后构建了一个基础模型在已有人工标注数据上进行训练，使得模型能够学习到隐性句逗号的标注逻辑和文本特征。接着，利用训练好的模型对未标注数据进行推断，生成伪标签。当模型输出的标签概率高于设定阈值时，将其作为新的标签数据，扩充至初始数据中。通过该方式，最终构建了一个包含33848条数据的隐性句逗号数据集，其中隐性句号19408条，隐性逗号14440条。识别实验中，将该数据集按照8:1:1的比例切分为训练集、验证集和测试集。具体统计结果如表 3所示。

**识别模型**：本文采用基于预训练语言模型的隐性句逗号分类识别方法。为充分考虑隐性句逗识别中相关语段特征，具体采用Hierarchical BERT模型(Lukasik et al., 2020)对句子对进行编码。如图 1所示，其包含了嵌入层、Sentence Encoder、Context Encoder层，模型输出为隐性句逗号的预测结果。其中Sentence Encoder使用了BERT预训练模型(Devlin et al., 2018)，Context Encoder使用GRU模型。模型的输入为两段文本，分别表示为Sentence1和Sentence2，此处并不是直接拼接两个句子，而是将其层次化并行输入模型。随后Sentence Encoder将学习每个Sentence句内的局部特征并聚合至对应的[CLS]向量中；将两个Sentence的[CLS]表征输入到Context Encoder层中，该层即可赋予文本前后的顺序特征以及


(c) 2024 中国中文信息学会计算语言学专业委员会

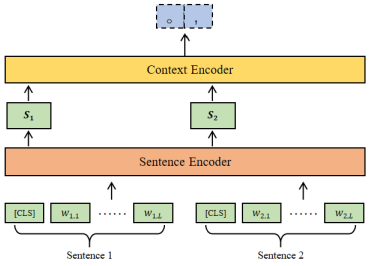



图 1: 隐性句逗号识别模型

| 模型 | accuracy(%) | precision(%) | recall(%) | f1(%) |
|---|---|---|---|---|
| Hierarchical BERT | 97.90 | 97.82 | 97.92 | 97.86 |
| Cross-segment BERT | 93.10 | 92.94 | 93.05 | 93.00 |
| BERT-CRF | 79.34 | 96.14 | 79.34 | 86.94 |
| BERT-LSTM-CRF | 86.17 | 97.58 | 86.17 | 91.52 |

表 4: 不同模型在隐性句逗号识别上的性能

上下文关系特征，并最终得到文本对的全局特征。全局特征通过线性分类层，将输出映射到隐性句逗的类别标签上，其中线性分类层通过Softmax函数对输出的概率进行了归一化，并利用交叉熵损失函数计算损失。

**识别实验**：除基于Hierarchical BERT模型外，我们也进行了基于Cross-segmen BERT模型(Wicks and Post, 2021)的方法。与此同时，我们还对比实现了基于序列标注的方法，包括BERT-CRF(Liu et al., 2020) 、BERT-LSTM-CRF(Yang et al., 2022)。实验表明，基于Hierarchical BERT的方法性能最佳。各模型结果见表 4。原因在于Hierarchical BERT相比可以充分学习到句逗号相关语段文本的词汇、语序及上下文特征。而序列标注方法的问题在于，当前的隐性句逗号识别任务下相关数据是独立的文本段，并非实际文本中的句逗号序列。基于Hierarchical BERT的隐性句逗号识别模型将用于后续机器翻译汉语长句化短的预处理阶段。

## 4 基于隐性句逗号识别的长句机器翻译

### 4.1 本文方法模型

为验证基于隐性句逗号识别的长句机器翻译方案效果，使用PipeLine方案 (Atrio et al., 2023)进行实现。首先，对源语汉语进行预处理，即进行隐性句逗号识别，并与源语文本比对校正，确定最终句逗号；然后对预处理文本进行机器翻译。如图 2。

**隐性句逗号识别**：模型所输入的文本为一个句逗号切分的标点句序列（记为sentence1,2,.....n)。为适配本文基于Hierarchical BERT的隐性句逗号识别模型，相邻的两个标点句组合作为一个输入，经过识别模型，预测其间的标点位置为隐性句号或隐性逗号。

**句逗校正**：由于隐性句逗号识别模型仅考虑了隐性句逗号，而实际文本中为所有句逗号（既包括隐性句逗号，也包括真句逗号），这里需对隐性句逗号模型识别结果进行校正,以获得最终句逗标点。具体做法为，将隐性句逗识别模型输出结果与原文结果进行比对，当模型输出结果与原文一致，保留原文本标点；当模型输出结果与原文本不一致，保留隐性句逗的识别结果。这样做的原因在于，由于受预训练语言BERT自身所包含的大量一般句逗号文本知识的影响，隐性句逗号识别模型并不能完全准确地从真实文本（包含所有句逗号）中识别出隐性句逗号。具体而言，当其标点分类结果与原文本不一致时，可以认为是，句逗模型增强了本文隐性句逗号知识后的结果，即为隐性句逗号；当其与原文一致，可以认为是BERT自身包含的大量真句逗号知识的结果。校正后的句逗标点文本，作为预处理结果输入机器翻译模型。


(c) 2024 中国中文信息学会计算语言学专业委员会

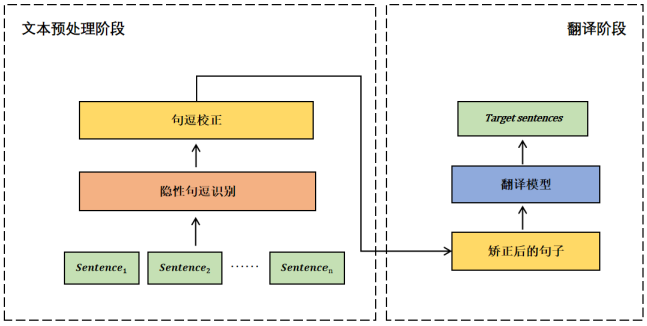



图 2: 基于隐性句逗号识别的机器翻译架构

| | 数据 | 文本长（逗号数） | 基准翻译 | 本文方法 |
|---|---|---|---|---|
| Opus-mt | WMT2018 新闻 | 3 个逗号以下 | 17.12 | 16.43 |
| | | 包含3 个逗号 | 15.88 | 16.74 |
| | | 包含4 个逗号 | 15.14 | 16.11 |
| | | 包含5 个逗号 | 15.30 | 16.58 |
| | WMT2023 文学 | 3 个逗号以上 | 16.42 | 17.11 |
| Randeng-mt | WMT2018 新闻 | 3 个逗号以下 | 18.42 | 17.93 |
| | | 包含3 个逗号 | 15.88 | 16.75 |
| | | 包含4 个逗号 | 14.65 | 15.24 |
| | | 包含5 个逗号 | 13.93 | 14.39 |
| | WMT2023 文学 | 3 个逗号以上 | 14.29 | 14.18 |

表 5: 翻译结果比较

## 4.2 实验设置

**数据集**：为验证本文方案效果，本文在标准机器翻译任务WMT2018的中英平行翻译数据集（新闻）[1]和WMT2023中的中英文学翻译平行语料[2]的测试集上进行基于隐性句逗号识别的长句机器翻译实验。引入文学语料是考虑，文学文本是机器翻译难点，而且文学文本相比一般文本口语性强，可逗可句或可断可连的情况更突出，特别适于本文方法的验证。为了充分验证本文方法在汉语长句上的效果，进一步按照文本长度（以包含的逗号数计算）将数据集划分如表 3。考虑是，逗号是句内标点，一个文本包含的逗号数越多，往往意味着其中的句子包含的小句越多，句子越长。

**翻译模型**：在翻译阶段，本文使用两个表现优异的预训练机器翻译模型进行了实验，分别是opus-mt (Tiedemann and Thottingal, 2020)和Randeng-mt (Zhang et al., 2022)。二者均基于transformer的端到端架构，并基于大型中英平行语料库训练，包含了丰富的双语知识。实验中二者具体设置相同，主要是：设置束搜索宽度为2；采用多概率采样；top_k为40，其可以在生成过程中过滤掉不太可能的词，仅保留头部的40个词；top_p概率为0.8，其代表输出词的概率之和达到或超过0.8，才会在生成过程中保留；此外为确保模型不重复地连续输出，限制了模型连续重复生成单词的个数为5。

**评估指标**：采用通用的BLEU值 (Papineni et al., 2002)。其主要通过比较机器翻译的输出和参考译文的n-gram 相似度计算得出。BLEU值中的n-gram的取值为4。

[1]https://www.statmt.org/wmt2018/translation-task.html
[2]https://aclanthology.org/events/wmt-2023/2023wmt-1


(c) 2024 中国中文信息学会计算语言学专业委员会

| 样例 | 类别 | 文本内容 |
|---|---|---|
| 1 | 原文 | **S1**四是织密织牢民生保障网，**S2**增进人民福祉。**S3**我们坚持以人为本，**S4**持续增加民生投入[，] **S5**保基本、兜底线、建机制，**S6**财政用于民生的比例达到70%以上 |
| | 原文翻译 | Four are the well-being of the people through the woven network of secure livelihoods. In order to promote the well-being, we insist on people-centred and sustained increases in human investment[，] and the preservation of basic, bottom-line and institutional mechanisms, the share of the financial resources spent on human life is over 70 per cent. |
| | 修正文本 | **S1**四是织密织牢民生保障网，**S2**增进人民福祉。**S3**我们坚持以人为本，**S4**持续增加民生投入[。] **S5**保基本、兜底线、建机制，**S6**财政用于民生的比例达到70%以上 |
| | 修正文本翻译 | The fourth is to weave a network of secure livelihoods for the betterment of the people. To improve the well-being of the people, we insist on people-centred efforts to increase people's inputs[.] To protect the basics, to get the bottom line, to build the mechanisms, the proportion of the funds spent on people's lives is more than 70%. |
| 2 | 原文 | **S1**而此时还处在大厅里的人，**S2**都并不急于领取任务[，] **S3**他们都认为顾飞一定会来自首[。] **S4**而自首的地方同样是这个大厅，**S5**因此这些人都紧盯着进进出出的玩家[，] **S6**一发现是法师就两眼放光，**S7**接着又是一脸失望。 |
| | 原文翻译 | Many of the people inside the hall did not immediately receive the mission[ ] as they were confident that Gu Fei would turn himself in [. ] They closely observed all the players coming and going since this hall was also a place where fugitives could turn themselves in [. ] Their eyes would periodically gleam whenever they saw a Mage and then be followed by a look of disappointment. |
| | 修正文本 | **S1**而此时还处在大厅里的人，**S2**都并不急于领取任务[。] **S3**他们都认为顾飞一定会来自首[，] **S4**而自首的地方同样是这个大厅，**S5**因此这些人都紧盯着进进出出的玩家[。] **S6**一发现是法师就两眼放光，**S7**接着又是一脸失望。 |
| | 修正文本翻译 | And no one in the hall is in a hurry to get a job [. ] They all think that Luo Fei will turn himself in[, ] and that the same hall is where they turn themselves in, so they're all staring at the players who come in and out [. ] The wizards are looking at the light, and then they're disappointed. |

表 6: 原文及译文与修正文本及译文对比

### 4.3 实验结果

如表 5所示：

（1）随着逗号增多，即句子包含的小句数增多，句长增大，机器翻译效果变差，充分证明长句越长对机器翻译的挑战越大。

（2）在包含3到5个逗号的语段文本内，本文方法比基准方案的翻译质量整体有所提高，其中在更稳健翻译系统（Opus-mt）上，本文方法的BLEU值分别提高了0.86、0.97和1.28，也即随着源语文本长度的增加，本文方案的优势变得更明显。

（3）本文方法也适应于文学翻译，在Opus-mt系统上，本文方法比基准模型提升了0.69个BLEU值。一般认为文学翻译难度大，主要是文学文本中更多人物对话和叙事，也更多涉及日常生活，内容容易理解，但也因此句子口语性强，结构更灵活，可断可连的句子更多，句子边界相比更模糊，翻译断句难度更大。本文方法可为文学机器翻译难题的解决提供一种特别思路。

（4）值得注意，本文方法对于相对较短的句子（包含3个以下逗号的语段文本）翻译，并没有体现出优势。这一方面反映出短句翻译并非机器翻译难题，另一方面也反映出本文方法也还未能更好识别真句逗号。这是因为隐性句逗号识别模型仅专注于隐性句逗号的识别，而暂时忽视了真句逗号的问题。这无疑需要在进一步工作中予以合理解决。

### 4.4 实例分析

表 5分别给出了汉语原文与其机器翻译结果和经本文方法修正标点后的汉语修正文本与其机器翻译结果。

**实例1** 文本选自WMT2018的新闻文本。修正文本将S4后的逗号改为了句号，修正后S3、S4、S5、S6间的关系远近更清晰，逻辑语义更明确。对比修正文本翻译与原文翻译可以发现，修正文本翻译更好地反映了S3与S4的关系，S5与S6的关系，而且相比原文翻译语法


(c) 2024 中国中文信息学会计算语言学专业委员会

结构也更合理。

**实例2** 文本选自WMT2023的文学文本。修正文本将S2后的逗号改为了句号，将S3后的句号改为了逗号，将S5后的逗号改为了句号。修正后S3、S4、S5间的紧密关系得以凸显，包括S3与S4的递进性关系（都是“他们都认为”的内容，并且用“而”连接），S5与S3、S4间的因果性关系（由“因此”体现）。并S3-5与S1-2和S6-6的关系距离也相对拉开，显得更清晰。对比修正文本翻译与原文翻译，修正文本翻译较好地反映了各个S间关系的语义远近与逻辑关系关系。

## 5 总结

针对长句机器翻译难题，本文提出基于隐性句号逗识别的汉语长句机器翻译方法。本文首先构建了隐性句逗号号数据集，并基于预训练模型实现了隐性句逗自动识别；进而将隐性句逗号识别模型作为翻译数据预处理模块与翻译模型结合，以解决汉语长句翻译难题。实验结果显示本文方法对长句的翻译性能有显著提升，而且长句越长效果越佳；这一方法也提升了文学翻译的长句翻译性能。未来我们将进一步完善隐性句逗号识别方法，特别是完全考虑真实语境中所有句逗号（真实句逗号与隐性句逗号）的识别问题，并将探索直接在机器翻译模型中融入隐性句逗号的理解。

## 参考文献


Àlex R Atrio, Alexis Allemann, Ljiljana Dolamic, and Andrei Popescu-Belis. 2023. A simplified training pipeline for low-resource and unsupervised machine translation. In *Proceedings of the The Sixth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2023)*, pages 47–58.

Junxuan Chen, Xiang Li, Jiarui Zhang, Chulun Zhou, Jianwei Cui, Bin Wang, and Jinsong Su. 2020. Modeling discourse structure for document-level neural machine translation. *arXiv preprint arXiv:2006.04721*.

Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:181004805*.

Chooi-Ling Goh and Eiichiro Sumita. 2011. Splitting long input sentences for phrase-based statistical machine translation. In *The Association for Natural Language Processing*, pages 802–805.

Jianming Guo, Xinran Chen, Zihan Liu, Weijie Yuan, Jianshen Zhang, and Gongshen Liu. 2022. Context modeling with hierarchical shallow attention structure for document-level nmt. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Bui Thanh Hung, Nguyen Le Minh, and Akira Shimazu. 2012. Sentence splitting for vietnamese-english machine translation. In *Proceedings of the 2012 Fourth International Conference on Knowledge and Systems Engineering*, pages 156–160.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *First Workshop on Neural Machine Translation*, pages 28–39.

Shaohui Kuang and Deyi Xiong. 2016. Automatic long sentence segmentation for neural machine translation. In *Natural Language Understanding and Intelligent Applications: 5th CCF Conference on Natural Language Processing and Chinese Computing, NLPCC 2016, and 24th International Conference on Computer Processing of Oriental Languages, ICCPOL 2016, Kunming, China, December 2–6, 2016, Proceedings 24*, pages 162–174. Springer.

Daniel Li, I Te, Naveen Arivazhagan, Colin Cherry, and Dirk Padfield. 2021. Sentence boundary augmentation for neural machine translation robustness. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7553–7557. IEEE.

Mingyi Liu, Zhiying Tu, Zhongjie Wang, and Xiaofei Xu. 2020. Ltp: a new active learning strategy for bert-crf based named entity recognition. *arXiv preprint arXiv:2001.02524*.

Michal Lukasik, Boris Dadachev, Kishore Papineni, and Gonçalo Simões. 2020. Text segmentation by cross segment attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.



(c) 2024 中国中文信息学会计算语言学专业委员会

Francisco Oliveira, Fai Wong, and Iok-Sai Hong. 2010. Systematic processing of long sentences in rule based portuguese-chinese machine translation. In *Computational Linguistics and Intelligent Text Processing: 11th International Conference, CICLing 2010*, pages 417–426.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Jean Pouget-Abadie, Dzmitry Bahdanau, Bart van Merrienboer, Kyunghyun Cho, and Yoshua Bengio. 2014. Overcoming the curse of sentence length for neural machine translation using automatic segmentation. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 78–85.

Xin Tan, Longyin Zhang, Deyi Xiong, and Guodong Zhou. 2019. Hierarchical modeling of global context for document-level neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1576–1585.

Jörg Tiedemann and Santhosh Thottingal. 2020. Opus-mt–building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480.

Ha Nguyen Tien and Huyen Nguyen Thi Minh. 2019. Long sentence preprocessing in neural machine translation. In *2019 IEEE-RIVF International Conference on Computing and Communication Technologies (RIVF)*, pages 1–6. IEEE.

Rachel Wicks and Matt Post. 2021. A unified approach to sentence segmentation of punctuated text in many languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3995–4007.

Hao Xiong, Wenwen Xu, Haitao Mi, Yang Liu, and Qun Liu. 2009. Sub-sentence division for tree-based machine translation. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 137–140.

Chongluo Yang, Long Sheng, Zhongcheng Wei, and Wei Wang. 2022. Chinese named entity recognition of epidemiological investigation of information on covid-19 based on bert. *Ieee Access*, 10:104156–104168.

Bao sheng Yin, Junjun Zuo, and Na Ye. 2012. Long sentence partitioning using top-down analysis for machine translation. In *Proceedings of IEEE CCIS2012*, pages 1425–1429.

Jiaxing Zhang, Ruyi Gan, Junjie Wang, Yuxiang Zhang, Lin Zhang, Ping Yang, Xinyu Gao, Ziwei Wu, Xiaoqun Dong, Junqing He, Jianheng Zhuo, Qi Yang, Yongfeng Huang, Xiayu Li, Yanghan Wu, Junyu Lu, Xinyu Zhu, Weifeng Chen, Ting Han, Kunhao Pan, Rui Wang, Hao Wang, Xiaojun Wu, Zhongshen Zeng, and Chongpei Chen. 2022. Fengshenbang 1.0: Being the foundation of chinese cognitive intelligence. *arXiv:2209.02970*.

Zhuocheng Zhang, Shuhao Gu, Min Zhang, and Yang Feng. 2023. Addressing the length bias challenge in document-level neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11545–11556.

冯文贺. 2019. 汉英篇章结构平行语料库构建与应用研究. 科学出版社.

吕叔湘. 1979. 汉语语法分析问题. 商务印书馆.

宋柔. 2022. 小句复合体的语法结构. 商务印书馆.

李幸and 宗成庆. 2006. 引入标点处理的层次化汉语长句句法分析方法. 中文信息学报, 20(4):8–15.

贾爱鑫, 李军辉, 贡正仙, and 张民. 2024. 融合目标端上下文的篇章神经机器翻译. 中文信息学报, 38(4):59–68.

黄河燕and 陈肇雄. 2002. 基于多策略分析的复杂长句翻译处理算法. 中文信息学报, 16(3):1–7.


(c) 2024 中国中文信息学会计算语言学专业委员会

# 基于知识蒸馏的低频词翻译优化策略

郭逸帆，眷红英✉，阎子悦，许鸿飞

[1]郑州大学计算机与人工智能学院，河南郑州
570714651@qq.com; iehyzan@zzu.edu.cn; 1716958920@qq.com; hfxunlp@foxmail.com

## 摘要

神经机器翻译通常需要大量的平行语料库才能达到良好的翻译效果。而在不同的平行语料库中，均存在词频分布不平衡的问题，这可能导致模型在学习过程中表现出不同的偏差。这些模型倾向于学习高频词汇，而忽略了低频词汇所携带的关键语义信息。忽略的这些低频词汇也包含重要的翻译信息，可能会对翻译质量产生不利影响。目前的方法通常是训练一个双语模型，然后根据频率为词汇分配不同的权重，通过增加低频词的权重来提高低频词的翻译效果。在本文中，我们的目标是提高那些有意义但频率相对较低的词汇的翻译效果。本文提出使用知识蒸馏的方法来提高低频词的翻译效果，训练在低频词上翻译效果更好的模型，将其作为教师模型指导学生模型学习低频词翻译。进而提出一个更加稳定的双教师蒸馏模型，进一步保证高频的性能，使得模型在多个任务上均获得了稳定的提升。本文的单教师蒸馏模型在英语→ 德语任务上相较于SOTA进一步取得了0.64的BLEU提升，双教师蒸馏模型在汉语→ 英语任务上相较于SOTA进一步取得了0.31的BLEU提升，在英语→ 德语、英语→ 捷克语和英语→ 法语的翻译任务上相较于基线低频词翻译效果，在保证高频词翻译效果不变化的前提下,分别取得了1.24、0.47、0.87的BLEU提升。

**关键词：** 知识蒸馏 ； 机器翻译 ； 低频词翻译 ； 双教师模型

# Knowledge Distillation-Based Optimization Strategy for Low-Frequency Word Translation in Neural Machine Translation

**Yifan Guo, Hongying Zan✉, Ziyue Yan, Hongfei Xu**

[1]School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou, Henan
570714651@qq.com; iehyzan@zzu.edu.cn; 1716958920@qq.com; hfxunlp@foxmail.com

## Abstract

Neural machine translation typically requires a large number of parallel corpora to achieve good translation results. In different parallel corpora, there is an imbalance in word frequency distribution, which may lead to different biases in the learning process of the model. These models tend to learn high-frequency vocabulary while ignoring the key semantic information carried by low-frequency vocabulary. The overlooked low-frequency vocabulary also contains important translation information, which may have an adverse impact on translation quality. The current method is usually to train a bilingual model and assign different weights to vocabulary based on frequency. By increasing the weight of low-frequency words, the translation effect of low-frequency words can be improved. In this article, our goal is to improve the translation performance of meaningful but relatively low-frequency vocabulary. We propose the use of

第二十三届中国计算语言学大会论文集，第206页-第216页，太原，中国，2024年7月25日至28日。
卷1：主会论文
(c) 2024 中国中文信息学会计算语言学专业委员会

knowledge distillation to improve the translation performance of low-frequency words, training model with better translation performance on low-frequency words, and using it as teacher model to guide student in learning low-frequency word translation. Furthermore, a more stable dual teacher distillation model is proposed to ensure high-frequency performance and achieve stable improvements in multiple tasks. The monolingual teacher distillation model in this study achieved a further improvement of 0.64 BLEU on the English-to-German translation task compared to the state-of-the-art (SOTA) model. The bilingual teacher distillation model, on the other hand, achieved a 0.31 BLEU improvement on the Chinese-to-English translation task compared to the SOTA model. In the translation tasks of English-to-German, English-to-Czech, and English-to-French, compared to the baseline low-frequency word translation approach, the proposed models achieved BLEU improvements of 1.24, 0.47, and 0.87, respectively, while ensuring that the translation performance of high-frequency words remained unchanged.

**Keywords:** Knowledge Distillation , Neural Machine Translation , Translation of low-frequency words , Dual Teacher Model

## 1 引言

神经机器翻译模型通常需要大量的平行语料(Kalchbrenner and Blunsom, 2013; Cho et al., 2014; Bahdanau et al., 2014; Sutskever et al., 2014; Gehring et al., 2017; Vaswani et al., 2017)，而在机器翻译任务中，翻译模型通常倾向于选择高频词而忽略低频词。在自然语言中，存在着词汇不平衡的情况，不同的单词可以表示相同的含义，更存在着熟词僻意的情况。例如：以汉语中的“然而”举例，在汉英翻译模型中，模型会更倾向于将其翻译为相对高频的“however”，而忽略“nevertheless”。Gu et al. (2020)指出高频词和低频词之间存在严重的不平衡现象，翻译模型很少有机会在训练过程中去学习低频的真实标签，而这些低频词往往携带着重要的语义信息，通常是一些特定领域、文学作品或方言中的专有词汇，能够准确地表达出某些特定的概念或情感。

在以往的工作中(Luong et al., 2015; Jean et al., 2015; Li et al., 2016; Pham et al., 2018)尝试通过维护短语表或低频词表来改进稀有词汇的翻译效果，一些工作在模型中增加了额外的组件(Gulçehre et al., 2016; Zhao et al., 2018)，但是这些带来了额外的训练复杂性和计算成本。一些基于BPE(Sennrich et al., 2016)或词段(Wu et al., 2016)的模型在一定程度上有效的缓解了标签分布不平衡的现象，并且成为多数翻译模型(Vaswani et al., 2017)的标准流程。但是无论多大的细分力度，它们仍然面临着标签分布不均匀的现象。现在多数的翻译模型大多为每一个词分配以相同的权重，并没有考虑到词频不同的情况。在这种情况下，低频词的训练机会相较于高频词小很多，模型在训练中会逐渐忽略低频词带来的损失，导致与低频词相关的参数无法得到有效的训练。

受益于Li et al. (2021)在图像分类中使用知识蒸馏解决长尾数据问题的启发，本文提出使用知识蒸馏的方法去优化低频词翻译效果。本文提出了一种仅使用现有数据的简单有效的方法用于提升低频词翻译效果，不需要引入额外的数据、人工成本。使用低频数据集训练教师模型，提升教师模型的低频词翻译效果，使用低频词翻译效果强的教师模型指导学生模型，学生模型在保留本身学习高频词的倾向的同时，由教师模型指导学生模型的低频词翻译效果。在此基础之上，提出使用双教师模型分别指导学生模型高频和低频词的学习，进一步保证学生模型中高频词的翻译效果，使得模型更具有稳定性。

本文任务的主要目标是：提升翻译模型中低频词的翻译效果，同时保证高频词的翻译效果不被破坏。本文的主要贡献如下：

- 本文使用知识蒸馏的方法解决低频词翻译效果不理想的问题，提出基于知识蒸馏的低频词翻译优化模型，减缓其倾向于高频的趋势，显著提升了低频词的翻译结果，同时保证了高频词翻译效果不被破坏。

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

- 本文在基于知识蒸馏的低频词翻译模型的基础上，提出双教师知识蒸馏模型，使用两个教师模型同时指导高、低频词的学习，保证了高频词的翻译性能，进一步提升了模型的稳定性。

- 基于知识蒸馏的低频词翻译优化模型在英语→德语任务SOTA的基础上进一步获得了0.64的BLEU提升；双教师蒸馏模型在汉语→英语任务上相较于SOTA取得了0.31的BLEU提升；相较于基线模型，在英语→德语、英语→法语任务上取得了1.24和0.87的BLEU提升，在英语→捷克语任务上获得了0.47的BLEU提升。

## 2 相关工作

### 2.1 机器翻译中的低频词翻译优化

在翻译任务中，常见的低频词类型有生僻词翻译、特殊俚语、专业术语等。低频词增加模型的词表多样性，会给模型带来很大的计算负担，翻译模型在处理较大词汇量时具有局限性。一些工作尝试维护短语表或回退词(Luong et al., 2015; Jean et al., 2015; Li et al., 2016)来解决大词汇问题。而现在的主流技术是使用基于子词的方法(Sennrich et al., 2016; Luong and Manning, 2016; Wu et al., 2016)，这种方法大大减少了词汇量并很好的解决了生僻词表示困难的问题。机器翻译的本质是一个分类任务，类别不平衡问题的主要解决方法有两种：基于数据(Baloch and Rafi, 2015; Sutskever et al., 2014)和基于算法(Zhou and Liu, 2005; Lin et al., 2017)的方法。其中基于数据的方法主要采用过采样和欠采样的方法，通过采样的方法解决类别不平衡的问题；基于算法的方法主要为不同的词分配以不同的训练策略。Jiang et al. (2019)提出一种线性加权的方式，根据词频为翻译任务中的词语分配以不同的权重，以此来解决低频词翻译不充分的问题。Gu et al. (2020)在此基础之上，进一步提出了两种权重分配的方法，通过卡方分布函数和幂函数的方式进一步优化的低频词的翻译质量，达到了近年来的SOTA水平。

### 2.2 知识蒸馏在机器翻译中的应用

知识蒸馏是近几年来解决各类迁移学习任务的热门方法，学生模型可以从教师模型中快速学习到知识的特性赋予了蒸馏方法很多的可能性。Zhuang and Tu (2023)使用知识蒸馏技术，通过对预训练模型的掩码语言模型进行微调，显著的提高了翻译模型的性能。Zhang et al. (2023)验证了通过知识蒸馏的方法可以从已经训练好的翻译模型中获取到知识并能将特定领域的知识转移到学生模型中。而在知识蒸馏中，并不是教师的知识越多越好，教师与学生模型之间的差异过大可能会损坏整体的性能(Wang et al., 2021)。为了解决这种问题，Huang et al. (2022)提出了一种新的蒸馏损失函数，使学生模型的学习目标由原先需要与教师模型完全一致，变为学习教师模型的软标签之间的相对关系，大大降低了蒸馏学习的难度。类别不平衡问题在各种任务中都可以观察到(Wei et al., 2013; Johnson and Khoshgoftaar, 2019)，在图像领域中，Li et al. (2021)使用知识蒸馏技术优化了长尾数据分类不均衡的问题，这为本文提出使用知识蒸馏技术优化低频词翻译提供了思路。本文使用知识蒸馏的方法，学习教师模型中特定的“低频（高频）词领域”的知识，用于解决词语分布不均匀的问题。

## 3 方法

本文的目标是提升模型在低频词上的翻译效果，同时保证高频词的翻译不受影响。使用在低频词上翻译效果更好的模型作为教师模型指导学生模型低频词的学习，同时学生模型保留模型本身学习高频词的倾向，学生模型在训练时就会减缓其倾向于高频的趋势。

在蒸馏模型中，首先需要训练一个在低频词上效果更好的模型作为教师模型。根据实验结果，教师模型要在低频词翻译上有显著提高且高频词不被严重破坏才能起到指导低频词有效学习的作用。基于知识蒸馏的低频词翻译模型即使用一个在低频词上翻译效果更好的教师模型去指导学生模型，但是实验过程中发现，随着数据集规模的增长，蒸馏模型指导学生模型学习低频词的效率会变得不稳定，甚至会破坏高频词的翻译效果。基于此本文提出双教师蒸馏模型，额外使用一个高频词教师模型去指导模型保障高频词的翻译效果，实验结果表明双教师模型在任意规模的数据集上均表现出稳定的增长。在蒸馏过程中，为了进一步保证高频词不被严重破坏，本文在进行蒸馏微调时对所使用的数据集进行了采样处理。


(c) 2024 中国中文信息学会计算语言学专业委员会

### 3.1 基于知识蒸馏的低频词翻译模型

基于知识蒸馏的低频词翻译模型的模型图如图 1所示，其中教师和学生模型的均参照Vaswani et al. (2017)中的实验设置。两个模型的输入序列相同，记为$X = (x_1, x_2, ..., x_n)$，真实标签记为$Y^* = (y_1, y_2, ..., y_m)$，输入经过Transformer编码器,输入序列均被转为向量序列$\mathbf{E}_x = [E_x[x_1], E_x[x_2], ..., E_x[x_n]]$,其中$E_x[x_i]$表示输入序列中第$i$个词的词向量编码与位置向量编码之和，经过编码器N层线性层后，将第N层的输出输入到解码器。解码器对编码器的输出执行多头注意力机制计算，解码器第N层的最终输出得到了目标隐藏层状态记为：$\mathbf{Y}_s = [[s_1], [s_2], ..., [s_m]]$，如下公式1-2所示。

$$\mathbf{E}_x = Embedding(X) + Embedding(S_{pos}) \tag{1}$$

$$\mathbf{Y}_s = Decoder(Encoder(\mathbf{E}_x)) \tag{2}$$

模型在得到解码器最后一层的隐藏层状态后，经过Softmax函数做归一化处理得到每一个预测类别的概率。然而，通过Softmax形成的概率分布通常是一个比较分散的软标签向量，最终的预测类别概率集合中的大多数类别的概率占比很低，学生模型很难学到这一部分知识。对此，利用超参数温度T对Softmax函数的结果进行平滑处理。最终模型一共可以得到三个输出，分别为经过温度T蒸馏得到教师模型和学生模型的软标签，分别记为$Out^{tea}_{soft}$、$Out^{stu}_{soft}$，以及学生模型对$\mathbf{Out}_T$取最大值得到的模型最终输出结果记为学生模型硬标签$Out^{stu}_{hard}$。

模型最终的损失函数根据公式4所示，由学生模型硬标签$Out^{stu}_{hard}$和真实标签$Y^*$计算模型的学生损失，使用教师模型软软标签$Out^{tea}_{soft}$和学生模型软标签$Out^{stu}_{soft}$计算蒸馏学习的损失。最终整个模型的损失由学生损失和蒸馏损失分配以不同的权重求和所得。

$$Out_T = Softmax(\frac{\mathbf{Y}_s}{T}) \tag{3}$$

$$Loss = \alpha * CrossEntropyLoss(Out^{stu}_{hard}, Y^*) + \gamma * DIST\_KD(Out^{stu}_{soft}, Out^{tea}_{soft}) \tag{4}$$

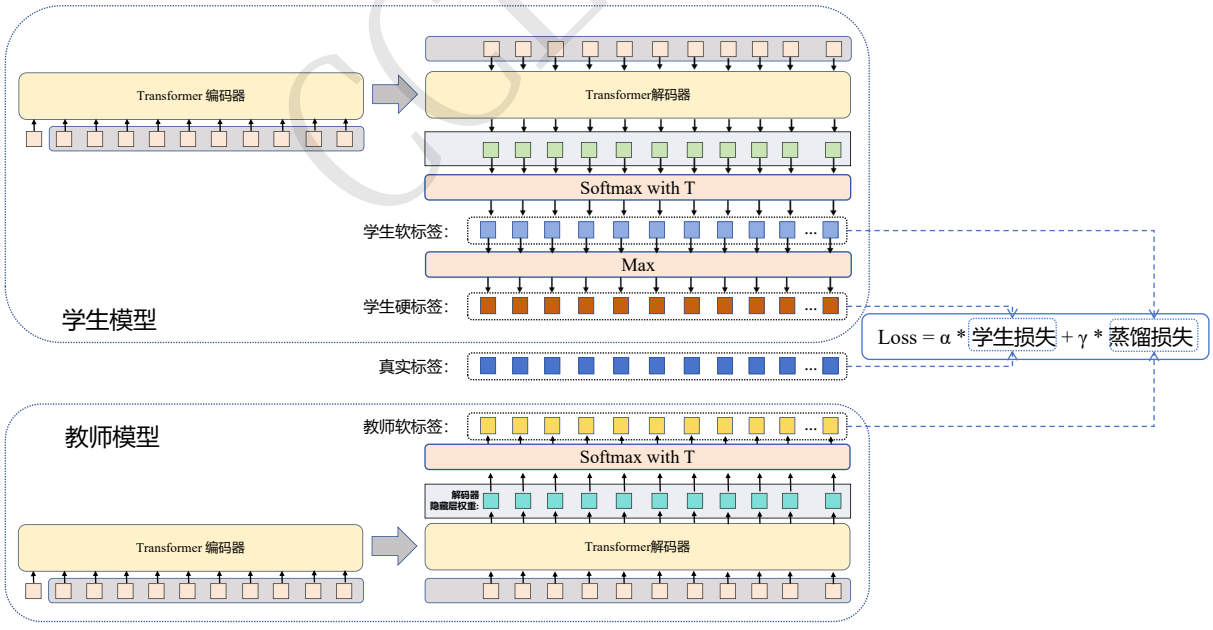



Figure 1: 基于知识蒸馏的低频词翻译优化模型


(c) 2024 中国中文信息学会计算语言学专业委员会

### 3.2 双教师知识蒸馏模型

随着数据规模的不断增大，数据集信息量增加而模型容量有限，低频教师模型指导学生模型学习低频词的程度进一步加深，低频蒸馏就会损伤模型的高频知识，从而导致高频词翻译效果不佳。具体而言，只用一个低频教师模型对小数据集的提升是有效的，但是在大数据集任务中表现不稳定。受到使用低频教师模型指导学生模型学习低频词的启发，在通过训练集采样限制模型高频词的学习倾向的同时，设置一个高频教师模型去限制学生模型高频词的大幅降低，将高频词的翻译效果限制在一个可控的范围之内，既不会受低频教师模型影响大幅降低，也不会过于倾向高频词。

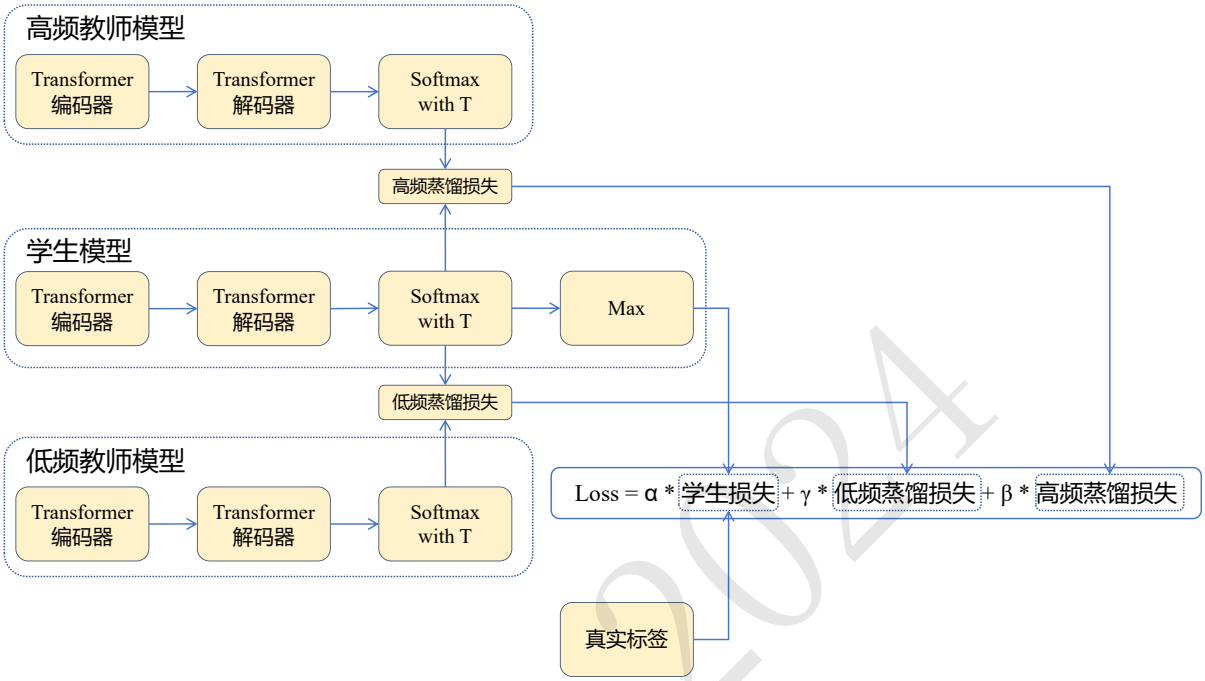



Figure 2: 双教师知识蒸馏教师模型指导学生模型

基于此，本文提出了双教师知识蒸馏模型，通过两个教师模型分别指导学生模型学习高低频词，双教师蒸馏模型的模型图如图2所示。双教师模型在蒸馏模型的基础之上，使用一个在高频词表现优秀的高频教师模型，并使用高频教师模型所得到的软标签$Out_{soft}^{h_tea}$与学生模型的软标签$Out_{soft}^{stu}$之间形成高频蒸馏损失，最终模型的整体损失如下公式5所示。

$$\begin{aligned} Loss = \alpha * CrossEntropy(Out_{hard}^{stu}, Y^*) + \\ \gamma * DIST\_KD(Out_{soft}^{stu}, Out_{soft}^{tea}) + \\ \beta * DIST\_KD(Out_{soft}^{stu}, Out_{soft}^{h\_tea}) \end{aligned} \tag{5}$$

## 4 实验

### 4.1 数据集

受Gu et al. (2020)启发，根据词频按照等式6将训练集和测试集进行打分，其中L表示句长，使用$\frac{1}{L}$消除句长影响；$y_i$表示句子中的每一个词在数据集中的频率，$y_k$表示整个词汇表中每一个词的频率。在等式6中，分母表示对整个词汇表中词频求和，在所有句子的得分计算中，分母均相同，句子中含有词汇的词频越小，分子越小，取对数后就会越小，取负号后就会越大，最终一个句子的得分越高表示该句子中包含的词相对低频。将训练集和测试集按照得分进行排序后将二者分别均分为三份数据集记为：$\{Train_{high}, Train_{middle}, Train_{low}\}$,$\{Test_{high}, Test_{middle}, Test_{low}\}$ 分别表示含有更多高频词的数据集、词分布均衡的数据集和含有更多低频词的数据集。蒸馏模型在进行知识蒸馏微调


(c) 2024 中国中文信息学会计算语言学专业委员会

| 数据集 | 训练集 | 验证集 | 测试集 |
|---|---|---|---|
| 英语->德语 | 4468668 | 3000 | 3003 |
| 英语->捷克语 | 10311551 | 3003 | 2656 |
| 汉语->英语 | 19489587 | 2002 | 2001 |
| 英语->法语 | 35705145 | 3000 | 3003 |

Table 1: 数据集统计表

时，使用采样的方法在三个子训练集中分别取出部分数据作为最终的微调数据集，即微调蒸馏模型的数据集为原数据集大小的1/3。

$$Freq_{sentence} = -\frac{1}{L}\sum_{i=0}^{L} log\frac{Count(y_i)}{\sum_{k=1}^{|V_t|} Count(y_k)} \tag{6}$$

本文主要使用了四个公开数据集，分别为WMT16的英语→德语数据集、WMT14的英语→法语数据集,两个数据集均分别选择newstest-2013和newstest-2014作为验证集和测试集；WMT18的汉语→英语数据集，使用该数据集在WMT官网提供的预处理版本，包含验证集与测试集；以及WMT15的英语→捷克语数据集，分别采用newstest-2013和newstest-2015作为验证集和测试集。按照以下步骤对数据集进行预处理：1）删除有错误编码的句子；2)将数据集中全角字符替换为相应的半角字符，并将所有的命名字符和数字字符（如>, >, &#x3e）转换为对应的Unicode字符；3）使用SentencePiece工具包(Kudo and Richardson, 2018)对数据集进行32000次联合BPE操作，以解决未知单词的问题;

### 4.2 蒸馏损失函数

Huang et al. (2022)证明在知识蒸馏中，更加强大的教师模型如果与学生模型差异比较大时，学生在普通的知识蒸馏模型上的表现可能会下降，甚至比在没有知识蒸馏的情况下从头开始训练还要糟糕。Huang et al. (2022)提出了一种只关注教师的偏好（即预测的结果相对排名）的方法。相较于传统的知识蒸馏损失关注恢复绝对值的做法，DIST_KD将学生模型从匹配一个强大的教师模型的确切输出的负担中解放出来，只关注于教师模型预测出的不同类别之间的相对关系。在本文中使用该方法，可以有效地减小学生模型对于教师模型的学习难度，模型的结果更加稳定。

### 4.3 教师模型

教师模型在选择上需优先选择在低频词上表现较好的模型，在Vaswani et al. (2017)的实验设置下，训练步数为100000得到基线模型记为Base，由于本文的方法都是在Base模型上进行蒸馏微调，所以我们额外在Base模型上不增加额外操作，以学习率为$1e-5$微调Base模型训练了一个BaseFT模型作为本文真正的基线模型，本文所有的实验结果均与BaseFT模型做对比。为了得到在低频词上表现更好的教师模型，在Base模型的基础之上，使用$Train_{low}$数据集在Base模型上进行微调，以此得到一个在低频词上表现效果更好的教师模型。在这种情况下得到的微调模型，在高频词的表现上大幅度降低，在$Test_{high}$上的结果低于Base模型。值得注意的是，DIST_KD在指导学生模型时依赖于教师模型预测结果的相对关系，所以低频教师模型在选择上需要保证BLEU结果在$\{Test_{high}, Test_{middle}, Test_{low}\}$上的结果递减。在高频词教师模型的选择上，因为BaseFT模型在微调时,依旧保留了其倾向于学习高频词的能力，所以BaseFT模型本身可作为一个在高频词上表现优异的教师模型。在双教师蒸馏模型中不同损失函数的权重上，为了保证学生损失的稳定，将$\alpha$置为1，而后以0.1为间隔在验证集上测试得到最优的损失函数权重$\alpha$、$\beta$、$\gamma$分别为1、0.6、0.4。

### 4.4 主要结果

本文的方法在WMT18中英数据集上的结果如表2所示，其中EMNLP_Exp与EMNLP_K2方法来自于(Gu et al., 2020)中的方法，代表当前的SOTA结果。在将测试集按照公式6划分为三个测试集$\{Test_{high}, Test_{middle}, Test_{low}\}$后，本文的方法在保证含有更多高频词的测试集$Test_{high}$的翻译效果不受影响的情况下，小幅度提升了词均衡的测试集$Test_{middle}$的翻译效


(c) 2024 中国中文信息学会计算语言学专业委员会

果，显著提升了含有更多低频词的$Test_{low}$的测试集的翻译效果。与当前的SOTA模型相比，本文的方法在高频测试集和均衡测试集上的效果基本与其持平或小幅提升。在低频测试集上进一步提升了0.31的BLEU值。相较于基线模型，整体提升了0.86的BLEU值，证明本文的方法十分具有竞争力。

| | 汉语→英语 | | | |
|---|---|---|---|---|
| | $Test_{low}$ | $Test_{middle}$ | $Test_{high}$ | Total |
| Base | 20.95 | 22.29 | 25.76 | 23.22 |
| BaseFT | 21.51 | 22.48 | 26.45 | 23.88 |
| $Teacher_{low}$ | 22.50 | 22.62 | 23.71 | 22.92 |
| EMNLP_Exp (Gu et al., 2020) | 22.06 | 22.57 | 26.54 | 24.01 |
| EMNLP_K2 (Gu et al., 2020) | 22.01 | 22.81 | 26.59 | 23.98 |
| Our_Double | **22.37** | 22.90 | 26.67 | 24.05 |

Table 2: 蒸馏模型在WMT18汉语→英语数据集上的结果

### 4.5 蒸馏模型在多种翻译任务上的结果

本文的两种方法在WMT16英德数据集上的结果如3所示，当前的SOTA模型相比，本文的方法在高频测试集和均衡测试集上的效果基本与其持平或小幅提升。在低频测试集上，单教师蒸馏模型的结果进一步提升了0.64的BLEU值，更加稳定的双教师蒸馏模型进一步提升了0.72的BLEU值。相较于基线模型，整体提升了1.24的BLEU值。

| | $Test_{low}$ | $Test_{middle}$ | $Test_{high}$ |
|---|---|---|---|
| BaseFT | 25.55 | 26.97 | 28.88 |
| EMNLP_Exp (Gu et al., 2020) | 26.07 | 27.33 | 28.91 |
| EMNLP_K2 (Gu et al., 2020) | 25.99 | 27.28 | 28.90 |
| Our_Sigle | 26.71 | 27.40 | 28.97 |
| Our_Double | **26.79** | 27.44 | 28.99 |

Table 3: 蒸馏模型在WMT16英语→德语数据集上的结果

为了测试方法在更多数据集上的可靠性与稳定性，分别选取了WMT15的英语→捷克语以及WMT14的英语→法语数据集继续验证本文的蒸馏方法。在WMT15的英语→捷克语和WMT14的英语→法语数据集上的实验结果如表4所示。在两个语种上的结果，双教师蒸馏模型相较于基线模型BaseFT,同样做到了保证高频词不被破坏且基本与基线模型持平，小幅度提升了均衡测试集结果，大幅度提升含有更多低频测试集的翻译效果。在英语→捷克语和英语→法语的低频测试集上分别提升了0.47和0.87的BLEU值，双教师蒸馏模型的稳定性得到了验证。

| | 英语→捷克语 | | | | 英语→法语 | | | |
|---|---|---|---|---|---|---|---|---|
| | $Test_{low}$ | $Test_{middle}$ | $Test_{high}$ | Total | $Test_{low}$ | $Test_{middle}$ | $Test_{high}$ | Total |
| Base | 27.74 | 27.98 | 30.39 | 28.48 | 37.16 | 39.58 | 41.47 | 39.65 |
| BaseFT | 27.80 | 28.42 | 31.40 | 29.07 | 37.75 | 40.16 | 41.93 | 40.23 |
| $Teacher_{low}$ | 28.82 | 28.08 | 29.09 | 28.64 | 38.67 | 40.21 | 40.22 | 39.76 |
| Our_Double | **28.27** | 28.64 | 31.49 | 29.33 | **38.62** | 40.45 | 42.01 | 40.65 |

Table 4: 蒸馏模型在英语→捷克语、英语→法语数据集上的结果

### 4.6 不同数量教师模型对蒸馏结果的影响

为了验证单教师模型和双教师模型在不同数据集上的影响，本文在每个数据集上均测试了单教师蒸馏模型和双教师蒸馏模型的实验效果。为了更好的表现数据集大小对不同方法稳定度的影响，选择了iwslt14的德语→英语的数据集补充为额外的小规模数据集。该数据集在数据量


(c) 2024 中国中文信息学会计算语言学专业委员会

上仅有174K句句对文件，相较于其它四个数据集，iwslt14数据集属于低资源数据，实验结果如表5所示。

实验结果表明，无论是单教师模型还是双教师模型在$Test_{low}$上的结果均表现出增长，证明使用蒸馏的方法对低频词的翻译效果是有效的。但是随着训练数据集规模的增长，单教师蒸馏模型在$Test_{high}$上逐渐显现出负增长，这是因为数据集信息量增加而模型容量有限，低频教师模型指导学生模型学习低频词的程度进一步加深，低频蒸馏就会损伤模型的高频知识，从而导致高频词翻译效果不佳。双教师蒸馏模型使用一个高频教师模型约束在蒸馏过程中对高频词的破坏，实验结果表明在任何规模的数据集上，双教师蒸馏模型均表现出稳定的增长。

| 任务 | 模型 | 测试集 | | | |
|---|---|---|---|---|---|
| | | $Test_{low}$ | $Test_{middle}$ | $Test_{high}$ | Total |
| iwslt14德语→英语 (174K句对) | BaseFT | 27.87 | 30.88 | 35.70 | 31.10 |
| | Single_Model | 28.73↑ | 31.54 | 35.67→ | 31.56 |
| | Double_Model | 28.74↑ | 31.59 | 35.77→ | 31.69 |
| WMT16英语→德语 (4.4M句对) | BaseFT | 25.55 | 26.97 | 28.88 | 27.21 |
| | Single_Model | 26.71↑ | 27.40 | 28.97→ | 27.88 |
| | Double_Model | 26.79↑ | 27.44 | 28.99→ | 27.93 |
| WMT15英语→捷克语 (10M句对) | BaseFT | 27.80 | 28.42 | 31.40 | 29.07 |
| | Single_Model | 28.58↑ | 28.06 | 29.59↓ | 28.68 |
| | Double_Model | 28.27↑ | 28.64 | 31.49→ | 29.33 |
| WMT18汉语→英语 (19M句对) | BaseFT | 21.51 | 22.48 | 26.45 | 23.88 |
| | Single_Model | 22.28↑ | 22.32 | 24.71↓ | 23.16 |
| | Double_Model | 22.37↑ | 22.90 | 26.67→ | 24.05 |
| WMT14英语→法语 (35M句对) | BaseFT | 37.75 | 40.16 | 41.93 | 40.23 |
| | Single_Model | 38.48↑ | 39.37 | 40.74↓ | 39.26 |
| | Double_Model | 38.62↑ | 40.45 | 42.01→ | 40.65 |

Table 5: 不同数量教师模型在不同规模数据集上的实验

### 4.7 不同损失函数对蒸馏结果的影响

本文在损失函数的选择上，选择了 Huang et al. (2022)提出的DIST_KD损失函数，而并非传统的KLD蒸馏损失函数。DIST_KD蒸馏损失函数相较于传统的KLD蒸馏损失函数，学生模型更加专注于学习教师模型预测的词之间的相对关系，即教师模型的翻译倾向，缓解因为教师模型与学生模型差距过大而导致的学习不充分的问题，并且在一定程度上降低了学习难度。传统的KLDloss在计算蒸馏损失时，关注于教师模型预测概率的绝对关系。而本文在蒸馏实验中，低频词模型和高频词模型与学生模型之间相差很大，我们希望模型能够学习到低频教师模型对低频词的倾向，而并非要求学生模型与教师模型完全一致，这样会破坏学生模型对于高频词的学习。DIST_KD契合本文使用教师模型指导学生模型的学习倾向的目的。以WMT18中英数据集为例，选用不同蒸馏损失函数的实验结果如表6所示，实验结果表明DIST_KD损失由于学习低频词预测的相对关系，降低了学习难度，不会严重破坏学生模型的高频词。传统的KLDloss由于学习教师模型的绝对关系，低频教师模型对学生模型造成的影响更大，促使学生模型以完全学习低频词为目标，破坏了高频词的翻译效果，从而在$Test_{high}$上表现不佳。使用DIST_KD不仅进一步提升了低频词的翻译效果，同时保证了高频词的翻译稳定性。

| | 汉语→英语 | | |
|---|---|---|---|
| | $Test_{low}$ | $Test_{middle}$ | $Test_{high}$ |
| BaseFT | 21.51 | 22.48 | 26.45 |
| +KLDloss | 21.83 | 22.85 | 26.29 |
| +DIST_KD | **22.37** | 22.90 | 26.67 |

Table 6: 不同蒸馏损失函数对蒸馏结果的影响


(c) 2024 中国中文信息学会计算语言学专业委员会

### 4.8 数据采样对蒸馏结果的影响

在进行限制高频词翻译效果下降的措施上，本文同时采用了采样方法，分别在三个训练集{$Train_{high}$,$Train_{middle}$,$Train_{low}$}上按照不同比例进行采样后形成新的训练集，为了排除最终结果的提升是否受采样方法的影响，在保证和蒸馏实验其它步骤一致的情况下，不使用蒸馏方法而选择采样训练集直接微调。以WMT18中英数据集使用采样数据集进行微调的实验结果如表7所示，实验结果表明仅采用采样数据进行微调时，虽然提升了低频词的翻译效果，但是对高频词造成了严重破坏，在$Test_{high}$上降低了0.79的BLEU值。而在采样的基础之上增加单教师模型，中英数据集由于数据规模性较大且低频教师模型进一步指导学生模型的低频词翻译效果，导致高频词效果进一步下降。而双教师模型通过高频教师模型的作用，保证了高频词的翻译效果。

| | 汉语→英语 | | |
|---|---|---|---|
| | $Test_{low}$ | $Test_{middle}$ | $Test_{high}$ |
| BaseFT | 21.51 | 22.48 | 26.45 |
| +sample data | 21.96 | 22.79 | 25.66 |
| +single_distill | 22.28 | 22.32 | 24.71 |
| +double_distill | **22.37** | 22.90 | 26.67 |

Table 7: 数据采样对蒸馏结果的影响

## 5 总结

在本文的研究中，我们关注到了机器翻译任务中类别不均衡的问题。类别不均衡会导致翻译模型倾向于翻译高频词，进而导致词汇多样性被破坏。为了提升那些低频但具有关键含义的词，本文研究了利用知识蒸馏技术解决低频词翻译不充分的问题，进而提出了两种基于知识蒸馏的低频词翻译优化模型。基于知识蒸馏的低频词翻译优化模型有效的提升了低频词的翻译质量，双教师蒸馏模型在此基础上进一步保证了高频词的效果，在保证高频词不被破坏的情况下显著提升了低频词的翻译效果，在多项翻译任务中均获得了稳定的增长。最后的结果表明，本文的方法可以实现显著的性能改善，相较于SOTA模型获得了进一步的提升。

## 致谢

本文受国家自然科学基金联合基金重点支持项目（U23A20316）资助。许鸿飞受国家自然科学基金青年项目（62306284）、中国博士后科学基金面上项目（2023M743189）和河南省自然科学基金青年项目（232300421386）资助，昝红英是本文的通讯作者。

## 参考文献

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Maher Baloch and Muhammad Rafi. An investigation on topic maps based document classification with unbalance classes. *Journal of Independent Studies and Research (JISR)*, 13(1), 2015.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. Convolutional sequence to sequence learning. In *International conference on machine learning*, pages 1243–1252. PMLR, 2017.

Shuhao Gu, Jinchao Zhang, Fandong Meng, Yang Feng, Wanying Xie, Jie Zhou, and Dong Yu. Token-level adaptive training for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1035–1046, 2020.


(c) 2024 中国中文信息学会计算语言学专业委员会

Çaglar Gulçehre, Sungjin Ahn, Ramesh Nallapati, Bowen Zhou, and Yoshua Bengio. Pointing the unknown words. In *Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics (ACL), 2016.

Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from a stronger teacher. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 33716–33727. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/da669dfd3c36c93905a17ddba01eef06-Paper-Conference.pdf.

Sebastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. On using very large target vocabulary for neural machine translation. In *53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL-IJCNLP 2015*, pages 1–10. Association for Computational Linguistics (ACL), 2015.

Shaojie Jiang, Pengjie Ren, Christof Monz, and Maarten de Rijke. Improving neural response diversity with frequency-aware cross-entropy loss. In *The World Wide Web Conference*, pages 2879–2885, 2019.

Justin M Johnson and Taghi M Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):1–54, 2019.

Nal Kalchbrenner and Phil Blunsom. Recurrent continuous translation models. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1700–1709, 2013.

Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, 2018.

Tianhao Li, Limin Wang, and Gangshan Wu. Self supervision to distillation for long-tailed visual recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 630–639, 2021.

Xiaoqing Li, Jiajun Zhang, and Chengqing Zong. Towards zero unknown word in neural machine translation. In *IJCAI*, pages 2852–2858, 2016.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

Minh-Thang Luong and Christopher D Manning. Achieving open vocabulary neural machine translation with hybrid word-character models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1054–1063, 2016.

Minh-Thang Luong, Ilya Sutskever, Quoc Le, Oriol Vinyals, and Wojciech Zaremba. Addressing the rare word problem in neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 11–19, 2015.

Ngoc-Quan Pham, Jan Niehues, and Alex Waibel. Towards one-shot learning for rare-word translation with external experts. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 100–109, 2018.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725. Association for Computational Linguistics (ACL), 2016.

CCL 2024

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Fusheng Wang, Jianhao Yan, Fandong Meng, and Jie Zhou. Selective knowledge distillation for neural machine translation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6456–6466, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.504. URL https://aclanthology.org/2021.acl-long.504.

Wei Wei, Jinjiu Li, Longbing Cao, Yuming Ou, and Jiahang Chen. Effective detection of sophisticated online banking fraud on extremely imbalanced data. *World Wide Web*, 16: 449–475, 2013.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

Yuanchi Zhang, Peng Li, Maosong Sun, and Yang Liu. Continual knowledge distillation for neural machine translation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7978–7996, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.443. URL https://aclanthology.org/2023.acl-long.443.

Yang Zhao, Jiajun Zhang, Zhongjun He, Chengqing Zong, and Hua Wu. Addressing troublesome words in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 391–400, 2018.

Zhi-Hua Zhou and Xu-Ying Liu. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on knowledge and data engineering*, 18 (1):63–77, 2005.

Yimeng Zhuang and Mei Tu. Pretrained bidirectional distillation for machine translation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1132–1145, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.63. URL https://aclanthology.org/2023.acl-long.63.

CCL 2024

# 融合确定性因子及区域密度的k-最近邻机器翻译方法

齐睿[1,2]，石响宇[1,2]，满志博[1,2]，徐金安[1,2]，陈钰枫[1,2*]
[1]北京交通大学，计算机科学与技术学院，北京，100044
[2]交通数据分析与挖掘北京市重点实验室，北京，100044
{20281284,22120416,zhiboman,jaxu,chenyf}@bjtu.edu.cn

## 摘要

k-最近邻机器翻译（kNN-MT）是近年来神经机器翻译领域的一个重要研究方向。此类方法可以在不更新机器翻译模型的情况下提高翻译质量，但训练数据中高低频单词的数量不均衡限制了模型效果，且固定的k值无法对处于不同密度分布的数据都产生良好的翻译结果。为此本文提出了一种创新的kNN-MT方法，引入确定性因子（CF）来降低数据不均衡对模型效果的影响，并根据测试点周边数据密度动态选择k值。在多领域德-英翻译数据集上，相比基线实验，本方法在四个领域上翻译效果均有提升，其中三个领域上提升超过1个BLEU，有效提高了神经机器翻译模型的翻译质量。

**关键词：** 机器翻译；K-最近邻检索；确定性因子

# A k-Nearest-Neighbor Machine Translation Method Combining Certainty Factor and Region Density

**Rui Qi[1,2],Xiangyu Shi[1,2],Zhibo Man[1,2],Jinan Xu[1,2],Yufeng Chen[1,2*]**
[1]Beijing Jiaotong University, School of Computer Science and Technology, Beijing, 100044
[2]Beijing Key Lab of Traffic Data Analysis and Mining, Beijing, 100044
{20281284,22120416,zhiboman,jaxu,chenyf}@bjtu.edu.cn

## Abstract

k-Nearest-Neighbor Machine Translation (kNN-MT) is an important research direction in the field of neural machine translation in recent years. It can improve the translation quality without updating the model, but there are still many problems to be solved: the number of each label in training data is not balanced, and the fixed k value can not produce good translation results for the data in different density regions. In this paper, we propose an innovative KNN-MT method to reduce the influence of imbalanced dataset by introducing a certainty factor (CF), and to optimize the search effect by adjusting the k value dynamically according to the data density around the test point. On the multi-domain German-English translation dataset, compared with the baseline experiment, our method has improved the translation performance in all four domains, three of which have improved by more than 1 BLEU.

**Keywords:** Machine Translation , K-Nearest-Neighbor Search , Certainty Factor

[1]陈钰枫（通讯作者）：chenyf@bjtu.edu.cn

第二十三届中国计算语言学大会论文集，第217页-第229页，太原，中国，2024年7月25日至28日。
卷1：主会论文
(c) 2024 中国中文信息学会计算语言学专业委员会

## 1 引言

随着深度学习的飞速发展，神经机器翻译已成为机器翻译任务中的主流方法，翻译效果和质量相比基于统计的机器翻译方法有了巨大飞跃。但目前的机器翻译技术仍然存在着局限性，比如模型的稀有词汇翻译能力不强，在处理不常见或特定领域的术语时，传统的模型往往表现不佳。为此，Khandelwal等人(2021)提出了k-最近邻机器翻译（kNN-MT），kNN-MT是一种基于单词级最近邻检索机制的神经机器翻译增强技术，无需进一步训练，就可以添加到任何预先训练过的神经机器翻译模型中，通过设计数据存储（Datastore）并在其中对相关的术语进行检索，来提高特定领域中单词翻译的准确率，是近年来神经机器翻译中的重要研究方向。

目前，kNN-MT的研究工作主要围绕以下三个方面展开：（1）检索加速化方法：通过减小数据存储体积、优化检索方式等角度提高检索效率。（2）精准检索提升方法：利用预训练模型、监督对比学习等方法对数据存储进行重构，或者调整表示空间使其更加有利于检索。（3）场景特定应用方法：利用kNN-MT模型的优势，提高跨领域、低资源等领域的翻译效果。然而，现有方法主要存在以下两方面问题：

- **数据存储中各分类数量不平衡问题**：传统的kNN算法检索过程中需要考虑到测试点周围各个单词的数量，导致检索生成目标语言单词的概率偏向出现频率高的常见词，使得模型在预测相对罕见单词（rare word）上的表现不佳。**因此，如何提高模型预测罕见单词的能力成为了挑战。**

- **表示向量空间分布不均匀问题**：在检索过程中，现有方法通过定义具体的“k值”来指定检索单词的个数。然而，在数据存储的不同区域内，如果数据密度不同，那么当前进行检索的k值也应当不同，此时的k值应该随着区域密度而产生变化。**因此，如何动态的选择k值成为了另一个需要解决的挑战。**

为了解决以上的挑战，本文提出一种融合确定性因子和区域密度的k-最近邻机器翻译方法（CF-kNN-MT）。具体地，我们引入CF计算模块和动态k值计算模块，前者根据检索到的单词在整个数据存储以及测试点周围区域出现的频率来对kNN概率分布进行调整，后者根据测试点周围区域密度划定一个范围，参考范围内的训练数据来动态确定一个适合测试点的k值。

综上所述，本文贡献总结为如下三点：

- 针对数据存储中高低频单词数量不平衡问题，我们引入确定性因子（CF）来对检索算法进行改进，结合不平衡学习生成新的kNN概率分布，提高了kNN算法修正标准NMT模型预测失误的能力。

- 针对表示向量空间分布不均匀的问题，我们设计了一种根据区域数据密度动态调整k值的算法，以减小各区域密度的差异对检索效果产生的影响。

- 在多领域德-英翻译数据集上，本文提出的模型在各个领域内效果均有提升，其中三个领域上的提升超过1个BLEU，提升较为明显。

## 2 相关工作

### 2.1 kNN-MT

k-最近邻机器翻译（kNN-MT）是在标准神经机器翻译的基础上对其结果进行kNN检索以增强其翻译效果的一种方法，其首先将训练数据的上下文表示和目标词作为键值对存储到一个大的数据存储（Datastore）中，然后从数据存储中检索离测试数据最近的k个可能的目标词，以帮助选择输出的下一个单词。

k-最近邻机器翻译(Khandelwal et al., 2021)的流程主要分为两步：数据存储的构建以及机器翻译模型的生成与检索。

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

**构建数据存储**：把训练集中源语言的句子作为NMT模型的输入，当输出第t个单词时，以模型的隐藏层表示$h_t$为键，输出标准答案为值，将构成的键值对储存起来，就构成了数据存储。

**翻译生成与检索**：将测试集的源语言句子作为NMT模型的输入，将输出的隐藏层表示作为查询向量，在数据存储中检索出离它距离最近的k个键值对，根据其值以及与查询向量的距离来构建一个kNN的概率分布，将其与NMT模型本身的概率分布线性结合在一起，利用构成的综合概率分布对输出的目标进行预测。

目前国内外针对kNN-MT的改进工作主要可以分为三类：研究检索加速化方法、研究精准检索提升方法，以及针对特定场景展开应用。

## 2.2 检索加速化方法

kNN-MT虽然大幅提高了翻译效果，但其翻译速度远不如标准的NMT模型。因此，许多学者致力于在保持其精度的同时提高kNN-MT的翻译效率，比如基于聚类来对数据存储进行降维和剪枝(Wang et al., 2022a)，既能细化检索结果，又能显著降低推理过程中的翻译延迟。或者为最近邻搜索构建一个更小的数据存储(Meng et al., 2022)，后面解码的步骤中只需要对该数据存储进行检索即可，提高了检索效率。也有学者对检索方式进行了改进，尝试从数据存储中检索整片单词块而不是单个单词(Martins et al., 2022)，或者从输入句子的邻居句子集的子集而不是所有句子中检索邻居目标单词(Deguchi et al., 2023)，使速度得到显著提升。

## 2.3 精准检索提升方法

同时，也有许多学者通过各种方法进一步提高kNN-MT翻译精度，以使其效果进一步加强。原始的kNN-MT直接使用解码器最后一层的隐状态来构建数据存储，这种方式仍有很大提升空间，因为k-最近邻检索是依据向量之间的相似度进行检索的，而数据存储的向量由标准NMT模型生成，只在机器翻译任务上进行过训练，并没有接受过相似度相关的训练，因此利用其构建的数据存储检索效果并不够优秀。针对该问题，Li等人(2022)提出利用预训练模型构建数据存储，使键相近的值也相近；Wang等人(2022b)则提出利用多个正样本和负样本的监督对比学习来把原来用于检索的隐状态向量h学习成有利于检索的向量z，把数据存储改造成利于检索的形式。而Zhu等人(2023)引入了少量的新参数用来对表示空间进行调整，将密切相关的向量拉进，使表示空间分布平滑，同时使用新的参数异步刷新整个数据存储，以异步获得新的kNN知识。还有一些学者试图对模型本身作出改进：Yang等人(2022)引入k个最近邻知识蒸馏（kNN-KD），训练阶段基础NMT模型直接学习kNN的知识，Jiang等人(2022)通过引入NMT置信度以及进行鲁棒训练来使模型具有具有更好的鲁棒性。

## 2.4 场景特定应用方法

还有一些学者试图将kNN-MT技术应用于特定场景。在跨领域翻译问题上，Cao等人(2023)利用语义相关键-查询对训练出来了一个修订器（reviser），利用其去修改数据存储中键的表示，实现对数据存储的重构，提升模型在下游领域上的翻译效果。而在仅有单语数据的领域内，Huang等人(2023)将kNN-MT引入其中，在正向和反向上使用两个预训练的泛域NMT模型，通过反复迭代实现了NMT的无监督域适应，在对应领域的效果良好。面对低资源领域翻译时，Liu等人(2023)利用kNN-MT进行迁移学习，在初始化、训练和推理全部阶段均转移知识，并选择性地从父数据存储中提取数据构造子感知数据存储（child-aware datastore），有效地降低了模型的过度置信度，同时提高了翻译精度。

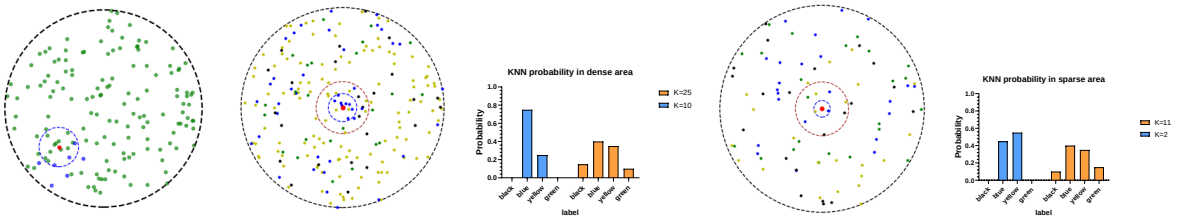
以上的三类方法为解决kNN-MT中的核心挑战提供了较好的思路，然而，这些方法仍存在一些局限性：（1）现有方法在提升翻译速度的同时无法保持良好的翻译效果，翻译效率和翻译质量达不到均衡。（2）利用机器翻译语料库构建的数据存储具有不平衡性、不均匀性等问题，均会影响kNN检索的效果。如何使数据存储和kNN检索更好地进行配合，这个问题仍然有待研究。因此，我们进一步提出了融合确定性因子及区域密度的k-最近邻机器翻译方法，主要解决了数据存储具有的不平衡性和不均匀性这两个问题。


(c) 2024 中国中文信息学会计算语言学专业委员会

# 3 模型介绍

## 3.1 问题分析

为了更好的理解本文所解决的问题，我们给出示例进行说明，具体如图1所示。




(a) 数据存储中各分类数量不平衡问题　(b1) 在数据密集区域内进行kNN检索　(b2) 密集区域概率分布示例　(c1) 在数据稀疏区域内进行kNN检索　(c2) 稀疏区域概率分布示例

Figure 1: kNN-MT方法中数据存储存在的问题示例

**（1）数据存储中高低频单词数量不平衡问题**

由于kNN是基于最邻近的k个样本来决定新样本的类别，如果某一类别的样本数量远多于其他类别，那么新样本更有可能被归为这个多数类别，从而导致分类偏差。我们假设图1a为一个数据存储某一区域的部分训练数据，其中绿色点对应的目标语言单词为A,蓝色点对应的目标单词为B。假设红色点是输入的测试数据，理想情况下应该将其翻译为单词B。我们在数据存储中对其进行kNN检索，可以看到，即使蓝色点基本都聚集在红色点周围，但由于绿色点在整个数据存储中数量都非常多，测试点还是会被分类为绿色点所代表的单词A。对于机器翻译领域，在任何语言中，负责连接和构建句子的功能词（如介词、冠词、连词等）往往出现得更频繁，而负责承载意义的内容词（如名词、动词、形容词等）出现的频率相对较低，这是因为不同的上下文中需要使用不同的内容词来表达特定的意思，但功能词往往可以通用。此外，人们在生活中不断创造新词汇和新表达方式，这些新词相比于传统表达方式出现的频率也更加罕见。因此，在kNN-MT领域针对数据不平衡问题开展研究，对于提高翻译效果至关重要。

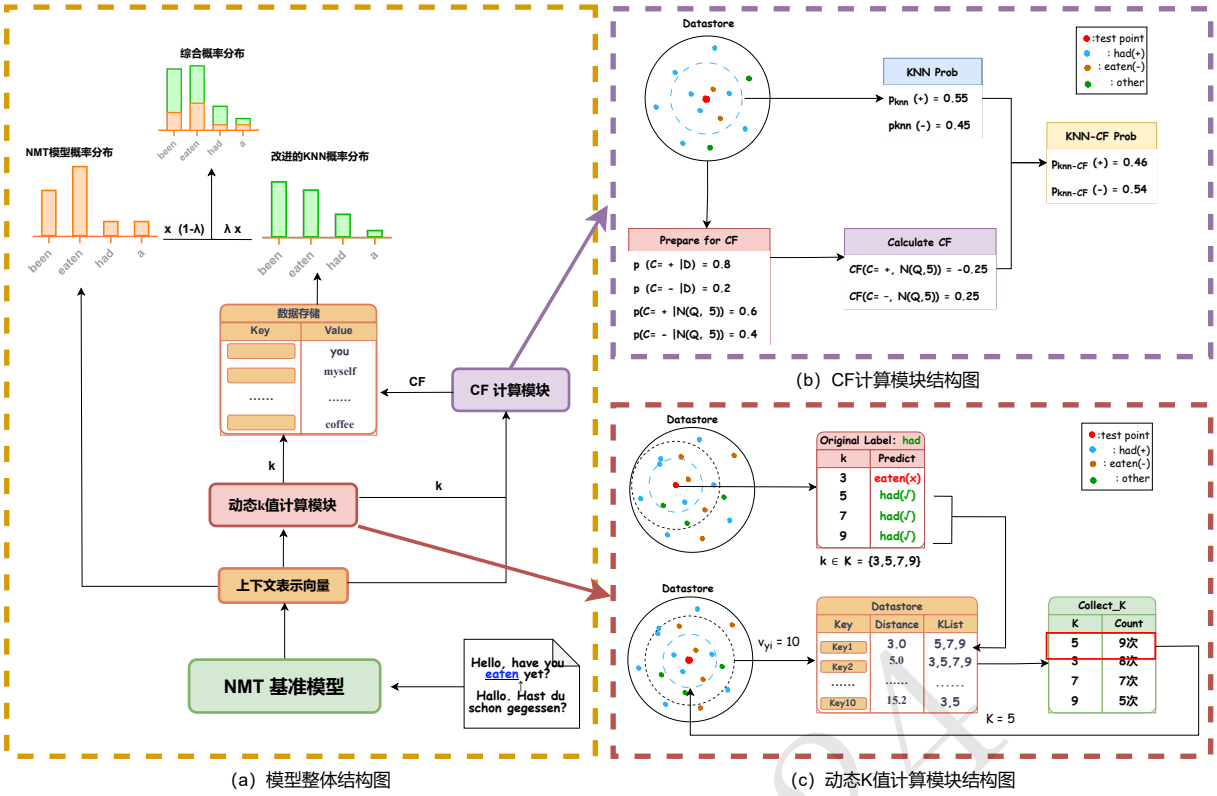
**（2）数据存储空间分布不均匀问题**

由于语言的多样性和复杂性，数据存储中用于检索的表示向量往往空间分布并不均匀，在不同区域呈现不一致的密度分布，这时使用固定的k值进行检索可能导致不准确的预测结果。假设图1b1和图1c1为一个数据存储的某两个部分，图1b1区域数据比较密集，图1c1区域则比较稀疏。如图1b1和图1b2所示，对于数据密集的区域，最近邻样本大量集中，较小的k值就足以反映出该区域的主要类别特征，较大的k值则容易引入噪声，影响估计的置信度。然而，对于数据稀疏的区域，如图1c1和图1c2所示，即使是最近的向量距离也较远，和测试点相似度不够高。如果取较小的k值（比如k=2），那么该点更有可能被分类为黄色，但由于测试点与检索到的两个最近邻点距离均比较远，特征不够相似，我们不能确信检索结果一定能够代表测试点的真实类别，这时就需要扩大检索的范围，通过检索周围更多数据来增强置信度。因此，我们有必要针对数据存储的空间分布不均匀问题展开研究，对k值进行合理规划。

因此，本文旨在通过解决数据存储中各单词数量不均衡与上下文表示空间密度分布不均匀的问题，提升k-最近邻机器翻译方法的翻译质量。

## 3.2 模型总体结构

我们对原始的kNN-MT模型进行了改进，整体结构如图2a所示。相比于原始模型，我们添加了CF计算模块，计算时考虑每个单词在整个数据存储中和在测试点周围的出现频率，进而优化kNN检索生成的概率分布；又添加了动态k值计算模块，参考和测试点处于相同密度区域的训练点来为每一个测试数据生成不同的k值，效果强于固定k值的方法。




(a) 模型整体结构图　(b) CF计算模块结构图　(c) 动态K值计算模块结构图

Figure 2: CF-kNN-MT模型结构

### 3.3 基于确定性因子（CF）测量的kNN分类

#### 3.3.1 确定性因子介绍

确定性因子模型最早由Shortliffe等人(1975)提出，是在基于规则的系统中管理不确定性的一种方法。确定性因子是分布于-1和1之间的数字，用于评估对一个规则的确定性或信心程度，-1表示完全不信任，0表示完全不确定，+1表示完全信任。0到1之间的值表示对命题或假设是真的有不同程度的信心，而0到-1之间的值表示对命题或假设是假的有不同程度的信心。如今，确定性因子在医学诊断、欺诈检测、客户服务、风险分析和自然语言处理等多个人工智能领域有实际应用，并已成功地用于识别数据集(Wu et al., 2004; Zhang and Wu, 2011)中的正关联规则和负关联规则。

#### 3.3.2 CF计算模块介绍

我们在kNN的概率计算中引入了CF，流程如图2b所示：假设D是整个数据存储，里面包含了若干个以上下文表示向量为键，对应单词为值的键值对；Q是测试点的表示向量，$N(Q,k)$是数据存储D中距离Q最近的k个向量所对应值的集合，$p(C=y_i|D)$是数据存储D中值为$y_i$的键值对数量与所有键值对数量的比值，$p(C=y_i|N(Q,k))$是值为$y_i$的键值对在集合$N(Q,k)$中的比值。那么我们可以利用上述符号来对CF进行表示。(Zhang, 2010)

如果$p\big(C=y_i|N(Q,k)\big) \geq p(C=y_i|D)$:

$$CF\big(C=y_i,N(Q,k)\big)=\frac{p\big(C=y_i|N(Q,k)\big)-p(C=y_i|D)}{1-p(C=y_i|D)} \tag{1}$$

否则:

$$CF\big(C=y_i,N(Q,k)\big)=\frac{p\big(C=y_i|N(Q,k)\big)-p(C=y_i|D)}{p(C=y_i|D)} \tag{2}$$


(c) 2024 中国中文信息学会计算语言学专业委员会

我们将CF与kNN概率进行线性组合，其中$p_{kNN}$表示KNN概率分布，$\omega$表示线性组合的插值参数：

$$p_{kNN-CF}(y_i|\mathrm{x},\,\hat{\mathrm{y}}_{<i}) = \omega\ CF\left(C = y_i, N\left(Q,k\right)\right) + (1-\omega)p_{kNN}(y_i|\mathrm{x},\,\hat{\mathrm{y}}_{<i}) \tag{3}$$

最后将$p_{kNN-CF}$与原始机器翻译模型生成的概率进行线性组合，就得到了模型输出第i个预测目标的概率分布，其中$p_{MT}$表示原始机器翻译模型的概率分布，$\lambda$表示线性组合的插值参数：

$$\mathrm{p}\ (y_i|\mathrm{x},\,\hat{\mathrm{y}}_{<i}) = \lambda\ p_{kNN-CF}(y_i|\mathrm{x},\,\hat{\mathrm{y}}_{<i}) + (1-\lambda)p_{\mathrm{MT}}\ (y_i|\mathrm{x},\,\hat{\mathrm{y}}_{<i}) \tag{4}$$

计算CF值的伪代码如代码1所示。

**Algorithm 1** Calculate CF

**Require:**
$y_i$：模型预测的第i个单词
$K_{num}$：在测试点周围选取的邻居数量（K）
$K_{dic}$：测试点周围K个邻居所对应单词的集合
$D_{num}$：整个数据存储的键值对数量
$D_{dic}$：整个数据存储中单词的集合
**function** CalculateProb:$(y_i, num, dic)$:
  $count = dic$集合中元素$y_i$的数量
  $p = count/num$
  **return** p

  $p1$ = CalculateProb$(y_i, K_{dic}, K_{num})$
  $p2$ =CalculateProb$(y_i, D_{dic}, D_{num})$
  **if** $p1 \geq p2$ **then**
    $CF = (p1 - p2)/(1 - p2)$
  **else**
    $CF = (p1 - p2)/p2$
  **end if**
  **return** CF

### 3.4 基于区域密度表示的动态k值算法

#### 3.4.1 局部密度与最近邻的关系

定理：(Mullick et al., 2018)假设一个d维超球面包含一个位于中心的点和内部一些均匀分布的点。当超球面包含的点数为$N_1$和$N_2$时，设距离中心最近的点与中心的期望距离分别为$d_{N1}$和$d_{N2}$。如果$N_1$>$N_2$ ，则$d_{N1}$<$d_{N2}$。

从定理可以看出，局部数据密度与最近邻距离近似成反比。因此，如果某一局部中某一数据的最近邻距离较大，我们可以假设该数据所在区域数据较为稀疏，而反之则表示区域数据比较密集。因此，我们使用最近邻距离作为查询点所在区域的密度表示，并按比例改变kNN-MT选取邻域的大小。

#### 3.4.2 动态k值计算模块介绍

为减小数据存储中各个区域的数据密度不均匀对结果造成的影响，我们设计了一种动态k值选取方法，尝试为每个查询点估计一个合适的k值，从而用动态的k值进行最近邻检索。

我们算法的核心思想如图2c所示：在预处理阶段将数据存储中的每一个键值对看做测试点，分别选取不同的k值，将该点在除本身以外的数据存储中进行检索，把能够正确分类该点的k值记录下来,形成集合KList,把每个键值对和对应的KList集合一同存储进数据存储中。在对测试数据进行检索时，我们参考每个测试点周围的$v_{y_i}$个邻居所适合的k值，选定检索该测试点所用的k值。确定k值的流程如下：计算CF值的伪代码如代码1所示。

CCL 2024

**Algorithm 2** Build KList

**Require:**
K=[3,5,7,9]:备选K集合
KList[key]:能够将键值对(key,value)正确分类的k值构成的集合
D:数据存储，里面存有许多个键值对(key,value)

```
for (key,value) in D do
  for k' in K do
    if 对key进行kNN检索(k=k')，检索出来的结果为value then
      k' −> KList[key]
    end if
  end for
end for
return KList
```

**（1）确定$v_{y_i}$:**

对于数据存储中任意一个键值对，我们假设键为$x_i$,我们将除它本身之外距离他最近的键值对与其的距离设为$d_{1NN}^{P/x_i}(x_i)$ ,即为最近邻距离。遍历数据存储中的所有键值对，计算每一个键值对的最近邻距离，其中最大的最近邻距离记为$d_{max}$，最小的最近邻距离记为$d_{min}$。即:

$$d_{max} = \max_{x_i \in P} d_{1NN}^{P \setminus \{x_i\}}(x_i) \tag{5}$$

$$d_{min} = \min_{x_i \in P} d_{1NN}^{P \setminus \{x_i\}}(x_i) \tag{6}$$

接下来，我们分三种情况讨论$v_{y_i}$:

$$v_{y_i} = \begin{cases} v_{\max}, & d_{1NN}^{P}(y_i) < d_{\min} \\ 1, & d_{1NN}^{P}(y_i) > d_{\max} \\ \left\lceil \left(v_{\text{lin}}\left(y_i\right) v_{\exp}\left(y_i\right)\right)^{0.5} / \delta \right\rceil, & d_{\min} \le d_{1NN}^{P}(y_i) \le d_{\max} \end{cases} \tag{7}$$

其中$\lceil . \rceil$符号代表向上取整, $\delta$ 是负责调整$v_{y_i}$大小的参数,$v_{\max} = \sqrt[3]{n}$。也就是说，若$y_i$周边数据足够密集，$v_{y_i}$被定为$v_{max}$；若$y_i$周边数据过于稀疏，$v_{y_i}$被定为1。在普遍情况下，$v_{y_i} \in [1,\ v_{max}]$。

其中，

$$v_{\text{lin}}(y_i) = [\frac{(1-\nu_{max})}{(d_{\max} - d_{\min})}(d_{1NN}^{P}(y_i) - d_{\min}) + \nu_{max}] \tag{8}$$

$$v_{exp}(y_i) = v_{max} e^{(d_{1nn}^{p}\ (y_i) - d_{min}\ )\beta} \tag{9}$$

$$\beta = -\frac{log_e v_{max}}{d_{max} - d_{min}} \tag{10}$$

在上述公式中，线性估计$v_{lin}\left(y_i\right)$将$v_{y_i}$建模为$d_{1NN}^{P}\left(yi\right)$在$v_{max}$和1之间的一个线性递减函数，对于较小的数据集能够维持一个可观的邻域大小；而指数估计$v_{exp}(y_i)$则是$d_{1NN}^{P}\left(y_i\right)$的指数递减函数，可以在较大的数据集中限制邻域不要过大。接下来，我们取这两种估计的几何平均值，因为相较于算术平均值而言，几何平均值更接近两个估计中较小的一个（即指数估计），该均值允许指数估计比线性指数估计对$v_{y_i}$的计算有更大的影响，这样的话即使对于较大的数据集，影响测试点的k值的邻域也可以保持较小。

**（2）k值选择:**

对于我们要进行检索的测试点$y_i$,假设上式计算出的$v_{y_i} = v$。对于一组函数$F = \{ f_1,\ f_2, \ldots\ f_{kmax} \}$，每一个$f_k \in F$ 都是从数据集X到{0,1}集合的多对一映射。其中$k_1, k_2 ..... k_{max} \in K$,$K$为可能会被用来进行kNN检索的k值的集合。

$$f_{k'}(x) = \begin{cases} 1\text{,如果}x \in X\text{且当}k = k'\text{时，}x\text{被}kNN\text{正确分类} \\ 0\text{,除此之外的情况} \end{cases} \tag{11}$$


(c) 2024 中国中文信息学会计算语言学专业委员会

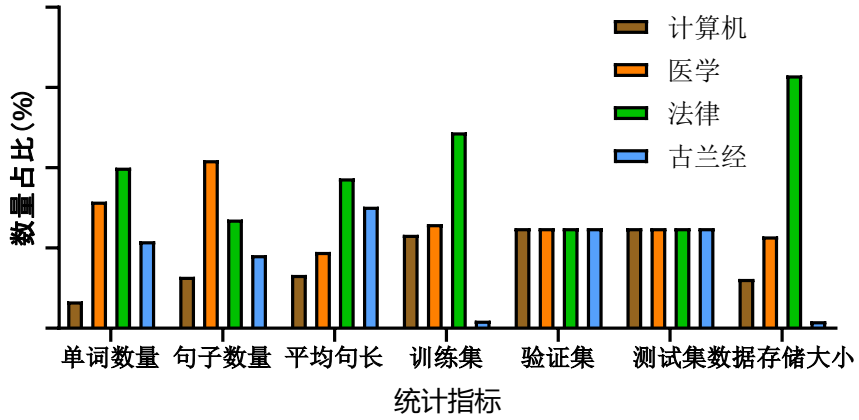



Figure 3: 不同领域数据集信息统计图

接下来我们定义:

$$z_{k'} = \sum_{x_i \in \tau_v^P(y_i)} f_{k'}(x_i) \tag{12}$$

$z_{k'}$表示当选取的k值为$k'$时，把数据存储中在$y_i$周围的每个键值对当做测试点，在数据存储中进行搜索，有多少个键值对可以被正确分类。因此，继续定义：

$$k_{yi} = \underset{k' \in \{1,2,..,k_{max}\}}{\text{argmax}} \quad z_{k'} \tag{13}$$

取得的$k_{yi}$即为用来对测试点$y_i$进行检索的k值。

## 4 评测实验与分析

### 4.1 实验设置

#### 4.1.1 实验数据集

为了验证该模型的效果，我们在几个常用的数据集上进行了实验：我们使用多领域德-英翻译数据集(Koehn and Knowles, 2017)进行德语-英语语言对实验，该数据集被广泛使用于机器翻译模型的效果评测中。我们跟随原始kNN-MT模型，在四个常用的基准上进行实验，其中的领域包括计算机、古兰经、医学和法律，以上数据集的细节如表1 所示。

| 数据集 | 计算机 | 医学 | 古兰经 | 法律 |
|---|---|---|---|---|
| 单词数量 | 3,041,677 | 14,301,472 | 9,848,539 | 18,128,173 |
| 句子数量 | 337,817 | 1,104,752 | 480,421 | 715,372 |
| 平均句子长度 | 9.0 | 12.9 | 20.5 | 25.3 |
| 训练集 | 223K | 248K | 18K | 467K |
| 验证集 | 2K | 2K | 2K | 2K |
| 测试集 | 2K | 2K | 2K | 2K |
| 数据存储大小 | 3.71M | 6.90M | 524K | 19.0M |

Table 1: 不同领域数据集信息及对应数据存储大小

我们对四个领域的数据集进行对比分析，从图3中我们可以分析得出，计算机领域数据集单词量较少，句子平均长度较短，反映了技术文档的直接性和命令式的表达方式；法律领域单词量多，且平均句长最长，显示了法律文本的复杂性和条款的详尽性。而医学领域句子数量最多，且长度较短，反映出医学文本对结构性和准确性的需求；古兰经领域的句子长度较长，文本特征更为复杂，反映了宗教文本的语言风格，增加了翻译的复杂性和挑战性。


(c) 2024 中国中文信息学会计算语言学专业委员会

#### 4.1.2 对比模型

**Base NMT**：我们的基准模型，是一个基于大型Transformer架构的编码器-解码器模型，使用Fairseq工具包实现。模型使用了WMT'19德英新闻翻译任务的双语数据以及反向翻译和人工翻译数据进行训练，然后在newstest数据集上进行微调。在数据处理阶段，作者使用Moses标记器对所有数据进行标准化和标记化处理，使用32K联合字节对编码（BPE）进行子词分割。该模型可以在fairseq的库中获取到，原论文及模型下载链接：https://github.com/facebookresearch/fairseq/tree/main/examples/translation。

**Vanilla kNN-MT**：我们的基线模型(Khandelwal et al., 2021)，本文工作主要是在此模型的基础上进行改进。模型利用多领域德-英数据集为每个领域专门构建数据存储，翻译过程中将Base-NMT模型的输出结果在域内数据构建的数据存储中进行kNN检索以增强其翻译效果。数据处理阶段同样需要对数据进行标准化和BPE（字节对编码）处理。模型具体流程介绍见2.1节，模型代码下载链接：https://github.com/urvashik/knnlm。

**N-Gram kNN-MT**：我们和Lv等人(2023)的模型结果进行比较，该方法在原始kNN-MT模型的基础上进行改进，将相邻的n-gram表示向量连接起来作为键，而对应目标单词的元组作为值。

#### 4.1.3 超参数及实验设置

本着控制变量原则，在会影响模型训练结果的超参数上，我们尽量保持统一。

**温度参数T**：用以防止模型对单个近邻给予过高的概率，进而防止模型过拟合到最相似的近邻点。大于1的温度参数会使分布变平滑，从而提高kNN-MT的性能。在古兰经领域我们取T = 100，其他领域均取T = 10。

**插值参数$\lambda$**：用以控制kNN检索概率和NMT预测概率线性组合的权重，在计算机领域我们取$\lambda = 0.7$，其他领域均取$\lambda = 0.8$。

**波数beam**：用以控制波束搜索的束的个数，我们取beam = 4。

上述参数我们都与原始kNN-MT模型保持一致。此外，本文模型中新加入了两个参数：

**调节参数$\delta$**：用于控制$v_{y_i}$的大小，它与数据存储的大小有关。

**新插值参数$\omega$**：用于将CF值与kNN概率进行线性组合。

**K集合**：在预处理过程中用于对数据存储每个点进行检索的k值的集合，后续对测试点进行检索时使用的是该集合中的其中一个k值。

新加入参数的具体数值如表2所示。

| 数据集 | $\delta$ | $\omega$ | K集合 |
|---|---|---|---|
| 医学 | 0.25 | 0.5 | 3,5,7,9,11 |
| IT | 0.25 | 0.6 | 3,5,7,8,9,11,13,15 |
| 法律 | 0.15 | 0.8 | 3,5,7,8,9,11 |
| 古兰经 | 0.25 | 0.4 | 7,11,15 |

Table 2: 本实验有关超参数设置

#### 4.1.4 评测指标

在本文实验中，所有的翻译结果都由SacreBLEU(Post, 2018)用区分大小写的去标记化BLEU进行评估。BLEU的主要原理是通过比较候选译文与参考译文中的n-gram的匹配程度来评估译文质量，匹配程度越高，我们认为译文的质量越优。BLEU评分得分范围从0到1，1表示完美的匹配。SacreBLEU是一个用于计算机器翻译评估指标BLEU分数的常用工具,它采用WMT（机器翻译会议）标准标记法，可以正确计算去标记化输出的分数.

BLEU可能无法完全捕捉翻译的流畅性和语言多样性，并且对于不同的参考翻译可能存在一定的敏感性，因此我们还采用COMET(Rei et al., 2020)指标作为补充度量。与传统的基于n-gram匹配的BLEU评分不同，COMET利用新的跨语言预训练模型（如XLM-RoBERTa），通过比较机器翻译输出、参考翻译和源文本的语义相似性来评估翻译质量，能够更全面地反映翻译的准确性、流畅性和忠实度。


(c) 2024 中国中文信息学会计算语言学专业委员会

#### 4.1.5 模型细节

我们以Zhu等人(2024)研发的kNN-BOX框架[0]为基础，使用Faiss[1]来进行向量检索。本文实验均在一个NVIDIA 2080 Ti GPU上进行。

### 4.2 实验结果与分析

#### 4.2.1 实验结果

我们分别将各个领域的训练集数据构建成数据存储，再用相应领域的测试集进行测试，结果如表4所示。表中标记为"-"的数据表明原作者未在文章中给出。

| 模型 | 计算机 | 法律 | 医学 | 古兰经 | 均值 |
|---|---|---|---|---|---|
| Base NMT | 38.35/39.21 | 45.48/57.52 | 40.06/46.93 | 16.26/-1.37 | 35.04/35.57 |
| Vanilla kNN-MT | 45.58/49.64 | 61.01/66.03 | 54.19/53.46 | 20.38 /-0.53 | 45.37/42.15 |
| N-Gram kNN-MT | 46.42/- | 61.98/- | 55.27/- | **21.39**/- | **46.26**/- |
| 本文模型 | **46.73/50.35** | **62.15/66.12** | **55.29/54.84** | 20.86/**0.94** | **46.26/43.06** |
| 模型提升 | +1.15/+0.71 | +1.14/+0.09 | +1.10/+1.38 | +0.48/+1.47 | +0.97/+0.91 |

Table 3: 在不同领域的德-英翻译测试集上的评测得分（BLEU/COMET）

从表中可以看出，与之前的研究相同(Khandelwal et al., 2021)，原始kNN-MT在所有数据集上的表现都显著优于基础NMT。此外，本文模型在全部领域的BLEU评分相较基线系统均有明显提升，BLEU分数平均值超过原始的kNN-MT达到了0.97点。而在计算机、法律、医学三个领域，本文模型提升超过1个BLEU，且提升效果优于对比模型N-Gram kNN-MT。此外，本文模型在所有领域的COMET评测得分也高于基准模型。综合考虑两个评测指标，我们发现本文模型在医学领域的提升效果最为明显，根据表1我们猜测这是因为医学领域单词数量较多，里面包含许多专业术语，这些术语属于低频词汇，我们的CF模块可以让翻译模型加强对这些低频词汇的关注。

#### 4.2.2 案例分析

我们在四个领域的测试数据中选取了一些典型的例子，分别输出原始模型和本文模型的翻译结果，进行对比分析，如图4所示。我们可以看出，本文模型的翻译结果相比原始模型更加贴近标准答案，在计算机和法律这两个领域的翻译结果更加精准，用词更加专业；在更要求严谨性的医学领域，我们的表述更加准确，且表达方式与源语言更加相似，不易产生歧义。古兰经领域数据集中包含着具有高度的专业性和语境依赖性的宗教文本，现有模型翻译效果均不够优秀，但我们的模型在用词的准确性和句意的表现上更胜一筹。

#### 4.2.3 消融实验

为了验证CF计算模块和动态k值计算模块对模型性能的贡献，我们进行了一系列消融实验，结果如表4所示。为了保证结果的可比性，除了被移除的模块，其他所有设置均保持不变。

从结果中我们可以看出，分别移除CF模块和动态k值计算模块，以及同时移除两个模块之后，模型的性能相比完整模型在各个领域数据集上的BLEU值均有降低，这表明上述两个模块对翻译准确性的提高均起到了重要作用。

#### 4.2.4 探究实验

为了使实验结果更加可靠，我们进一步研究了一些重要的超参数对模型的影响，以及一些变量的取值。

**关于$v_{y_i}$的取值** 在计算$v_{y_i}$的取值时，我们要考虑$y_i$周边区域的局部信息，比如数据密度以及邻域数据的类型分布，把它限制在一个合理的范围。如果周边数据较为密集，那么我们可以选取较大的$v_{y_i}$，这样选择k值的时候可以参考更多周边邻居的意见；反之如果周边数据较为

[0] https://github.com/NJUNLP/knn-box
[1] https://github.com/facebookresearch/faiss


(c) 2024 中国中文信息学会计算语言学专业委员会

| 计算机 | 医学 |
| --- | --- |
| [源语言]:Beschreibe die Merkmale der einzelnen Stufen<br>[标准答案]:Describe the **level characteristics**<br>[KNN-MT]:Describe the **properties of each level**<br>[CF-KNN-MT]:Describe the **level characteristics**<br><br>[源语言]:zur Auswahl einer Hintergrundfarbe für die Präsentation.<br>[标准答案]:**dialogue box, with which you can** select the background colour **of** the presentation.<br>[KNN-MT]:**to** select a background color **for** your presentation.<br>[CF-KNN-MT]:**dialogue box, with which you can** select a background color **for** the presentation. | [源语言]:Atmen Sie so weit wie möglich aus.<br>[标准答案]:Breathe out as far as **is comfortable**.<br>[KNN-MT]:Breathe out as far as **it will go**.<br>[CF-KNN-MT]:Breathe out as far as **is comfortable**.<br><br>[源语言]:Anwendung bei Kindern und Jugendlichen unter 18 Jahren.<br>[标准答案]:Use in children and adolescents **under the age of 18 years.**<br>[KNN-MT]:Use in children and adolescents **below 18 years of age.**<br>[CF-KNN-MT]:Use in children and adolescents **under the age of 18 years.** |
| **法律** | **古兰经** |
| [源语言]:Die Aufstellung muß ausserdem folgende Angaben enthalten<br>[标准答案]:The recapitulative statement **shall also set out**<br>[KNN-MT]:The recapitulative statement **must also include the following information:**<br>[CF-KNN-MT]:The recapitulative statement **shall also set out**<br><br>[源语言]:Keine weiteren Berichtigungen wurden beantragt oder für notwendig gehalten.<br>[标准答案]:No **further adjustment** was **claimed** or **considered** necessary.<br>[KNN-MT]:No further **corrections** were **requested** or **deemed** necessary.<br>[CF-KNN-MT]:No **additional** adjustments were **claimed** or **considered** necessary. | [源语言]:Meine Macht hat mich verlassen.<br>[标准答案]:Vanished has my power from me.（意译：**我的力量消失了**）<br>[KNN-MT]:I am not able to escape.（意译：**我无路可逃**）<br>[CF-KNN-MT]:My power has left me.（意译：**我的力量消失了**）<br><br>[源语言]:Gott hat sich Abraham ja zum Vertrauten genommen.<br>[标准答案]:**And** God **chose** Abraham as friend.<br>[KNN-MT]:**Verily,** God **has taken** Ibrahim as friend.<br>[CF-KNN-MT]:God **chose** Abraham as a friend. |

Figure 4: 对原始模型和本文模型翻译效果进行对比分析

| 模型 | 计算机 | 法律 | 医学 | 古兰经 | 均值 |
| --- | --- | --- | --- | --- | --- |
| CF-kNN-MT | 46.73 | 62.15 | 55.29 | 20.86 | 46.26 |
| -CF | 45.97 | 61.22 | 54.40 | 20.59 | 45.56 |
| - 动态k值 | 45.75 | 61.86 | 54.66 | 20.54 | 45.70 |
| -（CF+动态k值） | 45.88 | 61.01 | 54.19 | 20.38 | 45.37 |

Table 4: 在德-英翻译数据集上进行消融实验的BLEU得分

稀疏，那我们应该选取较小的$v_{y_i}$，以剔除掉距离过远的邻居的影响。在实验中，我们使用公式7来计算$v_{y_i}$的取值，与上述理论保持一致。

**关于K集合的选取** 在实验之初，我们选取了较为广泛使用的1,3,5,7,9作为备选k值(Mullick et al., 2018)，对其进行多种排列组合，将不同组合放入K集合中，在多个数据集上进行测试，发现K集合中的值不可以过少，因为这样会限制最后对测试数据检索时可供选择的k值，使动态k值的效果不够明显，极端情况下就变成了固定k值的检索。但备选k值并非越多越好，因为这可能会增加模型在决策时的不确定性，特别是某些较大的k值可能并不适合给定的数据分布，还会增加预处理的时间与占用的空间。根据我们在四个基准数据集上的评测结果，每个数据集上的K集合已经在表2中说明。

这里我们注意到一个问题：如果K集合中的某些k值能够正确分类的周边训练点数量一样且均为最多，那该选择谁作为检索测试点的k值呢？在这里，我们有六种策略：取最小的k值（方法A）、最大的k值（方法B）、随机选取一个k值（方法C），选择均值（方法D），选择中位数（方法E），以及和距离测试点最近的键值对选择同样的k值（方法F）。我们以古兰经数据集为例，针对四种策略进行了对照实验，结果如表5所示。表中结果表明，方法F的效果最为良好，且结果稳定，我们分析这是因为k值的选择受周围数据密度影响很大，而距离测试点最近的点和测试点周边的密度最为相似，因此其对k值的选择和测试点的选择也更为相似，所以要侧重考虑。

**探究确定性因子的影响** 在本文中，确定性因子被用来表示某一类别在整个数据存储的某一部分中出现的频率与在整个数据存储中出现的频率之比。$CF\big(C=y_i|N(Q,k)\big)$的取值为[-1,1]，如果CF大于0，那么我们认为将测试点的值预测为$y_i$的概率应该增加；CF小于0，我们认为预测为$y_i$的概率应该减小。CF等于0，我们认为目标点应该被预测为$C=y_i$的概率与正常kNN相同。在探究确定性因子要如何影响kNN检索结果时，通过反复比对，我们发现令kNN概率与确定性因子线性结合能够得到最好的效果。

**探究动态k值的影响** 如3.1中所描述的那样，当测试点位于密度不同的区域时，能够使检索


(c) 2024 中国中文信息学会计算语言学专业委员会

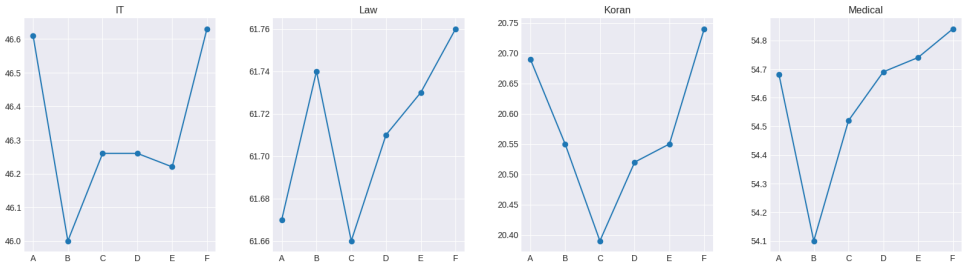



Figure 5: 面对多个k值均为最优时，不同方法选择对于模型效果的影响

效果达到最好的k值也不同，因此我们有必要根据测试点周围的密度动态选取k值。而在本文的方法中，k值的选择要参考测试点周围的数据，因为测试点周围的数据和测试点所处同样一片区域，周围的数据密度几乎一样，如果某一k值能够将测试点周边数据都正确分类，那么它也适合对测试点进行分类。

## 5 总结

本文针对k-最近邻机器翻译研究中存在的数据存储中高低频单词数量不平衡以及上下文表示空间密度分布不均匀两个尚未被关注的问题，提出了一种改进的神经机器翻译增强算法。在此过程中，我们取得了以下成果：

1.针对低频词较难被检索算法关注的问题，我们引入了确定性因子来增加这些数据被检索到的概率。确定性因子的引入在一定程度上平衡了各单词的数量不均衡问题，从而提高了kNN算法修正标准NMT模型预测失误的能力。

2.针对数据存储空间上密度分布不均匀的情况，我们设计了一种根据区域密度动态选取kNN检索所需的k值的算法。通过动态调整k值，我们能够减小不同区域数据密度差异对检索结果的影响，从而提高了检索的精度和可靠性。

在多领域德-英翻译数据集上，本文提出的改进神经机器翻译增强算法在各个领域内效果提升明显，均高于基准模型。未来我们将针对领域适应性开展研究，对该模型进行进一步改进，提升其在跨领域翻译等方面的性能。

## 致谢

本研究受国家自然科学基金面上项目(No.62376019,61976015, 61976016,61876198,61370130)资助。作者们还对匿名评审专家给予的宝贵建议表示衷心的感谢。

## 参考文献

Zhiwei Cao, Baosong Yang, Huan Lin, Suhang Wu, Xiangpeng Wei, Dayiheng Liu, Jun Xie, Min Zhang, and Jinsong Su. 2023. Bridging the domain gaps in context representations for k-nearest neighbor neural machine translation. pages 5841–5853. ACL.

Hiroyuki Deguchi, Taro Watanabe, Yusuke Matsui, Masao Utiyama, Hideki Tanaka, and Eiichiro Sumita. 2023. Subset retrieval nearest neighbor machine translation. pages 174–189.

Hui Huang, Shuangzhi Wu, Xinnian Liang, Zefan Zhou, Muyun Yang, and Tiejun Zhao. 2023. Iterative nearest neighbour machine translation for unsupervised domain adaptation. pages 13294–13301.

Hui Jiang, Ziyao Lu, Fandong Meng, Chulun Zhou, Jie Zhou, Degen Huang, and Jinsong Su. 2022. Towards robust k-nearest-neighbor machine translation. pages 5468–5477.

Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2021. Nearest neighbor machine translation.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In Thang Luong, Alexandra Birch, Graham Neubig, and Andrew Finch, editors, *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, August. Association for Computational Linguistics.

Jiahuan Li, Shanbo Cheng, Zewei Sun, Mingxuan Wang, and Shujian Huang. 2022. Better datastore, better translation: Generating datastores from pre-trained models for nearest neural machine translation. Number 1. Cornell University Library, arXiv.org, 2022-01-01. Report ER -.

Shudong Liu, Xuebo Liu, Derek F. Wong, Zhaocong Li, Wenxiang Jiao, Lidia S. Chao, and Min Zhang. 2023. knn-tl: k-nearest-neighbor transfer learning for low-resource neural machine translation. pages 1878–1891. ACL.

Rui Lv, Junliang Guo, Rui Wang, Xu Tan, Qi Liu, and Tao Qin. 2023. N-gram nearest neighbor machine translation, 01.

Pedro Henrique Martins, Zita Marinho, and Andr E. F. T. Martins. 2022. Chunk-based nearest neighbor machine translation. pages 4228–4245. EMNLP.

Yuxian Meng, Xiaoya Li, Xiayu Zheng, Fei Wu, Xiaofei Sun, Tianwei Zhang, and Jiwei Li. 2022. Fast nearest neighbor machine translation. pages 555–565. ACL.

Sankha Subhra Mullick, Shounak Datta, and Swagatam Das. 2018. Adaptive learning-based $k$ -nearest neighbor classifiers with resilience to class imbalance. *IEEE Transactions on Neural Networks and Learning Systems*, 29(11):5713–5725.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor, editors, *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium, October. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online, November. Association for Computational Linguistics.

Edward H. Shortliffe and Bruce G. Buchanan. 1975. A model of inexact reasoning in medicine. *Mathematical Biosciences*, 23(3):351–379.

Dexin Wang, Kai Fan, Boxing Chen, and Deyi Xiong. 2022a. Efficient cluster-based k-nearest-neighbor machine translation. pages 2175–2187. ACL.

Qiang Wang, Rongxiang Weng, and Ming Chen. 2022b. Learning decoupled retrieval representation for nearest neighbour neural machine translation.

Xindong Wu, Chengqi Zhang, and Shichao Zhang. 2004. Efficient mining of both positive and negative association rules. *ACM Trans. Inf. Syst.*, 22(3):381–405, jul.

Zhixian Yang, Renliang Sun, and Xiaojun Wan. 2022. Nearest neighbor knowledge distillation for neural machine translation. pages 5546–5556, July. Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies Carpuat, Marine de Marneffe, Marie-Catherine Meza Ruiz, Ivan Vladimir.

Shichao Zhang and Xindong Wu. 2011. Fundamentals of association rules in data mining and knowledge discovery. *Wiley Interdisc. Rew.: Data Mining and Knowledge Discovery*, 1:97–116, 03.

Shizhao Zhang. 2010. Knn-cf approach: Incorporating certainty factor to knn classification. *IEEE Intell. Informatics Bull.*, 11(1):24–33.

Wenhao Zhu, Jingjing Xu, Shujian Huang, Lingpeng Kong, and Jiajun Chen. 2023. Ink: Injecting knn knowledge in nearest neighbor machine translation. pages 15948–15959. ACL.

Wenhao Zhu, Qianfeng Zhao, Yunzhe Lv, Shujian Huang, Siheng Zhao, Sizhe Liu, and Jiajun Chen. 2024. knn-box: A unified framework for nearest neighbor generation. In Nikolaos Aletras and Orphée De Clercq, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024 - System Demonstrations, St. Julians, Malta, March 17-22, 2024*, pages 10–17. Association for Computational Linguistics.


(c) 2024 中国中文信息学会计算语言学专业委员会

# Ko-LLaMA:基于LLaMA的朝鲜语大语言模型

**庞杰[1,2,3]　闫晓东[1,2,3,*]　赵小兵[1,2,3]**
[1]中央民族大学　信息工程学院　北京　100081
[2]国家语言资源监测与研究民族语言中心
[3]民族语言智能与安全治理教育部重点实验室
*通讯作者：闫晓东
{22302041,yanxiaodong}@muc.edu.cn,nmzxb_cn@163.com

## 摘要

大语言模型在这两年受到了非常广泛的关注，像ChatGPT和GPT-4这样的大型语言模型（LLMs）极大地改变了自然语言处理研究，并在通向人工通用智能（AGI）的道路上迈出了令人兴奋的步伐。尽管已经开源了LLaMA等几个大型语言模型，但这些模型主要关注英文和中文语料库，对其他语言的适用性有限。而对于少数民族语言如朝鲜语来说，大语言模型的适用性更加有限。在本文中，我们通过扩展LLaMA现有的词表，增加了额外的20,000个朝鲜语Token，从而提高了其对朝鲜语的编码和语义理解的能力；并且进一步使用朝鲜语数据进行继续预训练，使用朝鲜语指令微调数据集对模型进行SFT（Supervised Fine-Tuning），并分析了不同数据量对指令精调效果的影响，经过继续预训练和指令微调后的模型显著提高了理解和遵循朝鲜语指令的能力。通过上述训练，极大增强了LLaMA的理解和生成朝鲜语文本的能力，并增强了其遵循指令的能力。实验结果表明，新提出的模型Ko-LLaMA显著提高了原版LLaMA在理解和生成朝鲜语内容方面的能力。此外，在鲜语文本分类数据集YNAT上对Ko-LLaMA与擅长少数民族语言的CINO模型及CINO的多种模型组合以及原版LLaMA和GPT-3.5进行了效果对比。结果表明，Ko-LLaMA的朝鲜语文本分类能力远超CINO和CINO的组合模型以及LLaMA和GPT-3.5等未经过朝鲜语语料进行词表扩充和继续预训练的大语言模型。

**关键词：** 朝鲜语；大语言模型；词表扩充；继续预训练；指令微调

# Ko-LLaMA: A Korean Large Language Model Based on LLaMA

**Jie Pang[1,2,3]　Xiaodong Yan[1,2,3,*]　Xiaobing Zhao[1,2,3]**
[1]School of Information Engineering, Minzu University of China, Beijing 100081
[2]National Language Resources Monitoring and Research Center for Minority Languages
[3]Key Laboratory of Ethnic Language Intelligent Analysis and Security Governance of MOE
Corresponding author: Xiaodong Yan
{22302041, yanxiaodong}@muc.edu.cn, nmzxb_cn@163.com

## Abstract

Large language models have gained immense popularity in the last couple of years, with models like ChatGPT and GPT-4 revolutionizing natural language processing research and taking exciting steps towards artificial general intelligence (AGI). Despite several large language models being open-sourced, such as LLaMA, these models primarily focus on English and Chinese corpora, with limited applicability to other languages. For minority languages such as Korean, the applicability of large language models is even more limited.In this paper, we enhance the applicability of LLaMA to the Korean language by extending its existing vocabulary with an additional 20,000 Korean tokens, improving its ability to encode and semantically understand Korean. We further


(c) 2024 中国中文信息学会计算语言学专业委员会

continue pre-training the model with Korean data, fine-tune the model with a Korean instruction dataset (SFT: Supervised Fine-Tuning), and analyze the impact of varying amounts of data on the fine-tuning effect. The model after continued pre-training and instruction fine-tuning significantly improves the model's ability to understand and execute Korean instructions. With the proposed approach, the capability of LLaMA to understand and generate Korean text is greatly enhanced, and its ability to follow instructions is strengthened. Experimental results show that the newly proposed model, Ko-LLaMA, significantly outperforms the original LLaMA in terms of understanding and generating Korean content. Furthermore, in the comparison of effectiveness on the YNAT dataset for fresh language text classification, Ko-LLaMA was compared against the CINO model, which excels in minority languages, along with various combinations of CINO models, original LLaMA, and GPT-3.5. The results indicate that Ko-LLaMA's ability in classifying Korean text far surpasses that of CINO and its combinations, as well as LLaMA and GPT-3.5, which have not undergone vocabulary expansion and continued pre-training on Korean language corpora.

**Keywords:** Korean , Large language model , Vocabulary extension , Continued pretraining , Command fine-tuning

## 1 引言

随着大型语言模型（LLMs）的出现，自然语言处理领域经历了实质性的范式转变。这些模型以其庞大的规模和全面的训练数据而受到关注，它们在理解和生成类似人类的文本方面表现出了非凡的能力。与专注于文本理解的预训练语言模型（如BERT）不同，GPT系列(Brown et al., 2020)强调了文本生成，使它们相比其他模型更适合发挥创造性。值得注意的是，GPT家族的最新成员，即ChatGPT和GPT-4，受到了极大的关注，他们在这个迅速发展的领域中确立了自己的领先地位。

然而，尽管LLMs具有很大影响力，但LLMs的实施具有很大的限制，这些限制阻碍了透明和开放的研究。最主要的问题是他们的专有性质，限制了对模型的访问，从而阻碍了更广泛的研究社区基于他们的成功进行建设。此外，训练和部署这些模型所需大量计算资源，对资源有限的大多数研究者来说是个挑战，进一步加剧了可研究性的问题。为了应对这些限制，NLP研究倾向于使用开源替代品，以增加更大的透明度和协作。LLaMA(Touvron et al., 2023a)、LLaMA-2(Touvron et al., 2023b)和Alpaca(San-Martin et al., 1968)就是这些倡议的显著例子。这些开源的LLMs旨在促进学术研究，加快NLP领域的进步。开源这些模型的目的是创造一个有利于模型开发、微调和评估的环境，最终创建适用于各种用途强大、有能力的LLMs，所以为了弥补少数民族语言朝鲜语在大语言模型上的空缺，我们研究一款可以用于少数民族朝鲜语的的大语言模型，便于以后朝鲜语大语言模型的研究与发展。

目前,中英文领域的大语言模型研究已经取得了很好的发展,而对于少数民族语言如朝鲜语来说,大语言模型仍处于萌芽阶段，没有发挥出应有的效果。中英文领域的大语言模型的词汇表几乎没有朝鲜语Token，无法编码和解码朝鲜语文本。通过对原版LLaMA模型进行词表扩充、继续预训练、指令微调等工作，极大的提高了LLaMA对朝鲜语的理解和生成的能力，充分挖掘了大语言模型在朝鲜语的能力。

本文的主要贡献如下：

1) 通过收集大量的朝鲜语训练语料，使用SentencePiece工具采用BPE的分词方式对语料进行分词，并与原版LLaMA词表进行合并，扩充了20,000个朝鲜语Token。并且采用低秩适配(LoRA)方法，对进行词表扩充后的模型进行了继续预训练，在使用较少计算资源的情况下显著提高了模型在朝鲜语的理解和生成能力。

2) 通过公开的YNAT朝鲜语文本分类数据集构造了包括45678条训练集、9104条验证集，9104条测试集的朝鲜语文本分类的指令微调数据集，并对模型进行指令微调，

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

分析了不同数据量对模型指令微调效果的影响，显著提高了模型朝鲜语文本分类的能力。通过在YNAT朝鲜语文本分类数据集上的对比实验，表明Ko-LLaMA的效果超过CINO及CINO相关组合模型。

## 2 相关工作

大语言模型在这两年取得了巨大的进步，接下来将现梳理大语言模型近几年的研究进展，再介绍词表扩充、继续预训练、指令微调一系列训练范式的发展状况。

OpenAI在2018年首次提出了GPT(Generative Pretrained Transformer)(Radford et al., 2018)。这种模型使用了Transformer的解码器架构，并使用了一个单向的语言模型目标进行预训练。在预训练后，GPT可以通过在特定任务的数据上进行微调来适应各种NLP任务，其工作原理与Bert类似，当作预训练模型使用。在2019年，OpenAI发布了GPT的第二版，即GPT-2(Radford et al., 2019)。与GPT相比，GPT-2有更多的参数（从1.1亿增加到3.4亿）。GPT-2在许多NLP基准测试中取得了领先的表现。2020年，OpenAI发布了GPT-3(Mann et al., 2020)，是当时最大的语言模型，拥有1750亿个参数。GPT-3在文本生成任务上的表现超越了许多先前的模型。在GPT-3的基础上，OpenAI进一步进行了微调，于2021年发布了InstructGPT(Ouyang et al., 2022)。这个模型经过了大量的模型训练和数据清洗，旨在理解和执行用户的指令。InstructGPT的训练过程包括两个阶段。第一阶段是预训练，这是在大量的互联网文本上进行的。而第二阶段是微调，这是在一个更小，特定的，由人类审核员生成的数据集上进行的。InstructGPT在许多实际应用中都表现出色，如进行技术支持、提供教育资源、帮助用户学习新技能等。由于InstructGPT的成功，OpenAI受到启发并于2022年11月底发布了ChatGPT，ChatGPT一经发布便受到了广泛的关注，迅速火遍全网。InstructGPT和ChatGPT都是从GPT-3微调而来，但他们的区别是，InstructGPT是为了理解和执行用户的指令，而ChatGPT则是为了进行自由形式的对话。基于上述区别，确定了朝鲜语大语言模型要先经过InstructGPT的训练阶段，但是由于GPT没有开源，但是上述工作要在开源的模型上展开，所以LLaMA(Touvron et al., 2023a)作为Meta开源的模型则是一个比较好的选择，而且LLaMA在各项NLP任务上取得了显著的成果。

LLaMA作为一款开源的大语言模型在大语言模型研究上做出了突出的贡献，由于LLaMA只在英文的效果上表现卓越，在其他语种上展现出来的能力并没有很突出，所以LLaMA在其他语言上的加固和拓展也基本采用了词表扩充+继续预训练+指令微调的训练范式。Chinese-LLaMA(Cui et al., 2024)是在LLaMA的基础上增强了中文能力，这也提供了一种思路。虽然加固其中文能力比开发出LLaMA的朝鲜语理解和生成的能力要简单一些，但是验证了这一思路的可行性。

## 3 Ko-LLaMA训练流程

### 3.1 原版LLaMA-7B词表扩充

LLaMA-7B的训练集大约包含1.4T的Token，其中大部分是英文，一小部分是拉丁语和其他欧洲语言(Xia et al., 2023)。因此，LLaMA几种语言上的能力表明了它具有多语种和跨语种理解能力。初步研究表明，LLaMA几乎没有朝鲜语理解和生成的能力，在朝鲜语的各类NLP任务上表现都有待提高。为了赋予LLaMA更强的朝鲜语理解和生成能力，我们使用朝鲜语语料库对LLaMA模型进行继续预训练。然而，直接使用朝鲜语语料库进行继续预训练面临一些挑战。首先，原始的LLaMA词汇表包含的朝鲜语Token非常少，这不足以编码一般的朝鲜语文本。虽然LLaMA Tokenizer可以通过回退到字节码来支持所有的朝鲜语字符，但这种回退策略会显著增加序列长度，因为每个朝鲜语字符都被拆分为3-4个字节Token，降低了朝鲜语文本的编码和解码效率。其次，字节Tokens不仅用于表示朝鲜语字符，还用于表示其它UTF-8 Tokens，所以使用字节Tokens的方法很难让LLaMA有效地学习捕获朝鲜语Token语义的表示。

为了解决这些问题并提高编码效率，我们使用额外的朝鲜语Token扩展LLaMA词表，并调整模型以适应扩展后的词表(Gao et al., 2023)。扩展过程如下：

1) 为了增强分词器对朝鲜语文本的支持，首先使用SentencePiece(Kudo and Richardson, 2018)采用BPE的方式在朝鲜语语料库上训练一个朝鲜语分词器，词汇量为20,000。


(c) 2024 中国中文信息学会计算语言学专业委员会

2) 随后，通过合并他们的词汇表，将朝鲜语分词器合并到原版LLaMA分词器中。这样就可以得到了一个合并过的分词器，就是包含了朝鲜语的LLaMA分词器，它的词汇量为49924。

3) 为了使LLaMA模型适应朝鲜语LLaMA分词器，我们将词嵌入层和语言模型头从V×H形状调整为V'×H，其中V=32000表示原始词汇表的大小，H表示词嵌入的维度，V' = 49924是朝鲜语LLaMA分词器的新词汇表大小。新添加的行被添加到原始嵌入矩阵的末尾，确保原始词汇表中的标记的嵌入不受影响。

| | Length | Content |
|---|---|---|
| 原始朝鲜语句子 | 80 | 최근 길림 연변변경관리지대 경신변경파출소 경찰은 영상순찰을 하던 중 야생동북범 한마리가 대도천촌 부근의 숲속에서 먹이를 찾는 모습을 발견했다. |
| 原版LLaMA Tokenizer | 137 | ['▁', '<0xEC>', '<0xB5>', '<0x9C>', '<0xEA>', '<0xB7>', '<0xBC>', '▁', '<0xEA>', '<0xB8>', '<0xB8>', '<0xEB>', '<0xA6>', '<0xBC>', '▁', '연', '<0xEB>', '<0xB3>', '<0x80>', '<0xEB>', '<0xB3>', '<0x80>', '경', '<0xEA>', '<0xB4>', '<0x80>', '리', '지', '대', '▁', '경', '신', '<0xEB>', '<0xB3>', '<0x80>', '경', '<0xED>', '<0x8C>', '<0x8C>', '<0xEC>', '<0xB6>', '<0x9C>', '소', '▁', '경', '<0xEC>', '<0xB0>', '<0xB0>', '은', '▁', '영', '상', '<0xEC>', '<0x88>', '<0x9C>', '<0xEC>', '<0xB0>', '<0xB0>', '을', '▁', '하', '<0xEB>', '<0x8D>', '<0x98>', '▁', '중', '▁', '<0xEC>', '<0x95>', '<0xBC>', '<0xEC>', '<0x83>', '<0x9D>', '동', '<0xEB>', '<0xB6>', '<0x81>', '<0xEB>', '<0xB2>', '<0x94>', '▁', '한', '마', '리', '가', '▁', '대', '도', '천', '<0xEC>', '<0xB4>', '<0x8C>', '▁', '부', '<0xEA>', '<0xB7>', '<0xBC>', '의', '▁', '<0xEC>', '<0x88>', '<0xB2>', '<0xEC>', '<0x86>', '<0x8D>', '에', '서', '▁', '<0xEB>', '<0xA8>', '<0xB9>', '이', '를', '▁', '<0xEC>', '<0xB0>', '<0xBE>', '는', '▁', '모', '<0xEC>', '<0x8A>', '<0xB5>', '을', '▁', '<0xEB>', '<0xB0>', '<0x9C>', '<0xEA>', '<0xB2>', '<0xAC>', '<0xED>', '<0x96>', '<0x88>', '다', '.', '▁'] |
| Ko-LLaMA Tokenizer | 35 | ['▁최근', '▁길림', '▁연변', '변경', '관리', '지대', '▁경신', '변경', '파출소', '▁경찰', '은', '▁영상', '순찰', '을', '▁하', '던', '▁중', '▁야생동', '북범', '▁한마', '리가', '▁대', '도천촌', '▁부근', '의', '▁', '숲', '속에', '서', '▁먹', '이를', '▁찾는', '▁모습을', '▁발견했다', '.'] |

图 1 原始LLaMA和Ko-LLaMA的tokenizer的效果比较

### 3.2 使用LoRA对LLaMA进行继续预训练

常规的大语言模型训练范式即更新LLMs的全量参数是非常昂贵的，并且在时间和成本上对大多数实验室或公司来说是不可行的。低秩适配（LoRA）(Hu et al., 2021)是一种参数高效的训练方法，它保持预训练模型的权重不变，同时引入可训练的秩分解矩阵。LoRA冻结预训练模型的权重，并在每一层注入可训练的低秩矩阵。这种方法显著减少了总的可训练参数，使得用更少的计算资源训练LLMs成为可能。具体来说，对于一个线性层，其权重矩阵$W_0 \in R^{d*k}$，其中k是输入维度，d是输出维度，LoRA添加了两个低秩分解的可训练矩阵$B \in R^{d*r}$和$A \in R^{r*K}$，其中r是预先设定的秩。带有输入x的前向传递公式如下:

$$h = W_0x + \triangle Wx = W_0x + BAx, B \in R^{d*r}, A \in R^{r*k} \tag{1}$$

在训练过程中，$W_0$是冻结的，不接收梯度更新，而B和A是可以更新的。通过选择秩rmin(d, k)，我们减少了内存消耗，因为我们不需要为大的冻结矩阵存储优化器状态。我们主要将LoRA适配器集成到注意力模块和MLP层的权重中，因为在QLoRA(Dettmers et al., 2023)中验证了将LoRA应用到所有线性Transformer块的有效性，表明这种选择是合理的。

我们使用标准的因果语言模型（CLM）任务对LLaMA进行朝鲜语上的继续预训练，给定输入序列$x = (x_0, x_1, x_2, ...)$，模型采用自回归的方式预测下一个Token，目标就是最小化负对数似然:

$$L_{CLM}(\Theta) = -\sum_i logp(x_i|x_0, x_1, , , , , x_{i-1}; \Theta) \tag{2}$$

经过继续预训练以后模型能够完成对已输入文本进行续写，即提高了模型对朝鲜语文本的理解和生成能力，接下来只需要对继续预训练以后的模型做任务上的指令微调就能让模型完成具体的下游任务。

### 3.3 朝鲜语指令微调

由于模型学习的知识是非常广泛的，所以我们对模型的输入输出难以控制，并且经常会生成一些无关的内容，这是因为公式(2)中的语言建模目标是预测下一个词而不是'follow instructions with human feedback.'(Ouyang et al., 2022)。为了使语言模型的行为与用户的意图相符，可以对模型进行微调，明确地训练它遵循指示。斯坦福Alpaca(Taori et al., 2023)是一个基于LLaMA的指令模型，它是在由Self-Instruct(Wang et al., 2023)中的技术生成的52K条指令数据上进行训练的。我们遵循斯坦福Alpaca的方法，对朝鲜语LLaMA进行有监督的微调训练，以训练出一个指令遵循模型。

由于数据有限，没有对继续预训练以后的LLaMA做大规模的指令微调，所以只挑选了一个NLP常见的任务：文本分类，挑选数据进行指令微调，并且与CINO(Yang et al., 2022)和CINO的一些组合模型以及原版LLamA和GPT-3.5等大语言模型做效果比对。

在进行指令微调数据之前，需要明确两个事情，一个是Prompt模版如何确定，另外一个是指令微调的数据从何而来。由于Chinese-LLaMA-Alpaca(Cui et al., 2023)已经成功进行了指令微调，所以这里的Prompt模版还是采用斯坦福Alpaca模版，不过因为我们是要做朝鲜语的指令微调，所以要将其格式转换成朝鲜语版本。具体Prompt形式如下图2所示：

| | |
|---|---|
| 아래는 작업을 설명하는 명령입니다.요청을 적절하게 완료하는 응답을 작성합니다.<br><br>###지시:<br>{instruction}<br><br>###반응:{output} | 下面是描述任务的指令。编写适当完成指令的回复。<br><br>###指令:<br>{instruction}<br><br>###回复:{output} |

图 2 指令微调的Prompt格式

通过使用上述Prompt模版进行构造数据，并且使用公开数据集YNAT经过数据扩充来构造指令微调的数据，YNAT公开的朝鲜语文本分类数据集，其中训练集45678条，验证集9104条，测试集9104条，并且将最终结果与CINO模型及CINO相关组合模型进行结果比较；同时为了验证与相关大模型在朝鲜语上理解能力的差距，设置了Ko-LLaMA与原版LLaMA和GPT-3.5在朝鲜语文本分类数据集上的对比实验。具体结果分析放在实验与分析环节。

## 4 实验与分析

### 4.1 继续预训练实验

#### 4.1.1 实验数据集

继续预训练的数据集主要是从新华网、人民网朝鲜语版等多家新闻媒体网站上爬取58642篇新闻文本作为训练集原始语料以及230篇新闻文本作为测试集原始语料。将部分语料爬取以后经过去除乱码、特殊符号、图片等信息过滤得到了比较纯净的新闻文本，共计680MB。

#### 4.1.2 模型评估标准

PPL（困惑度）是一种衡量语言模型质量的评价指标。它衡量模型预测下一个单词的难度。PPL 越低，模型越好。具体来说，PPL 是给定数据集上模型预测的单词序列的对数似然函数的负值。对于一个包含N 个单词的数据集，PPL 计算如下：

$$PPL = \exp\left(-\frac{1}{N}\sum_{i=1}^{N}\log P(w_i|w_1,...,w_{i-1})\right) \tag{3}$$

因为LLaMA在经过词表扩充和继续预训练以后，已经可以完成基本的文本续写，并且其训练预料大多是新闻文本，在朝鲜语新闻文本续写效果最为明显，所以对继续预训练以后的模型在230篇新闻文本上采用困惑度作为评估指标进行评估，这里需要注意一点，针对同一个测试集，采用不同tokenizer的模型会导致公式（3）中N的大小不同，即序列长度不同，所以tokenizer不同的模型进行PPL值的对比是没有意义的，因此本文采用词表扩充未经过继续预训练的模型和词表扩充并经过继续预训练的模型在230篇新闻文本续写任务的PPL值进行对比，这样保证tokenizer的一致性，也就是N相同，进而可以描述出模型经过继续预训练以后对朝鲜语的理解和生成能力的提升。

#### 4.1.3 实验设置

在继续预训练部分，由于数据集中没有其他大模型训练时用的那么多语料，并且原版LLaMA模型基本没有朝鲜语的编码和解码能力，所以我们设置的学习率等参数相对大一些便于模型快速学习朝鲜语语料的知识，下面表1为继续预训练的详细参数设置。


(c) 2024 中国中文信息学会计算语言学专业委员会

| 模型参数 | 参数值 |
|---|---|
| Peak learning rate | $2e-4$ |
| LoRA dropout | 0.05 |
| Batch size | 1024 |
| Max sequence len | 512 |
| LoRA rank | 8 |
| Torch dtype | Float 16 |
| LoRA alpha | 32 |
| Weight decay | 0.01 |

表 1 继续预训练的参数设置

在继续预训练阶段我们对模型进行了全面的预训练，包括Embedding层，这样不仅可以使模型适应新添加的朝鲜语Token，同时可以最小化对原始模型的干扰，并且我们向模型添加了LoRA权重（adapter），使用PEFT库去做带有LoRA的参数高效训练，同时使用DeepSpeed(Aminabadi et al., 2022)来优化训练过程中的内存效率，采用AdamW优化器(Loshchilov and Hutter, 2019)，峰值学习率为2e-4，warm-up cosine scheduler为5%。此外，我们还使用值为1.0的梯度裁剪，以缓解潜在的梯度爆炸。最终在两张A100 GPU（80GB VRAM）进行了一次5小时的迭代训练。

#### 4.1.4 实验结果分析

本节使用LLaMA-7B模型经过朝鲜语词表扩充以后做继续预训练，分别在爬取的训练数据上进行预训练，其中58642篇新闻文本按照9:1的比例划分训练集和验证集，测试集230篇新闻文本，验证集Loss下降图如图3所示。继续预训练阶段的Loss是前期急剧下降，然后变得平缓并趋于稳定，对此做出的分析是，在预训练阶段模型并不具备朝鲜语的编解码能力，模型参数还没有得到很好的调整，误差相对较大，因此每次参数更新都在较大程度上减小误差，所以损失下降得快；随着训练的进行，模型参数逐渐接近最优解，每次参数更新带来的误差减小，因此Loss下降速度变慢。

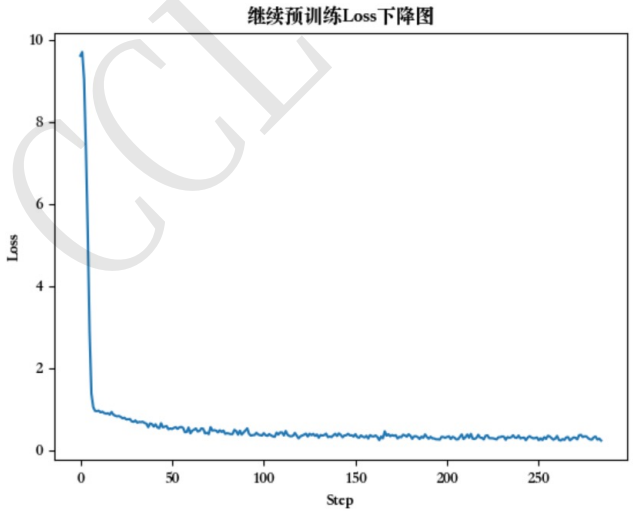



图 3 继续预训练Loss下降图

由于需要考虑PPL可比性的问题，所以在PPL对比实验中需要保证对比模型的词表一致，为了说明继续预训练对Ko-LLaMA的朝鲜语理解和生成能力的提升，我们将扩充词表但是没有经过继续预训练的模型和扩充词表并经过继续预训练的模型在230篇新闻文本上的续写任务计算得到的PPL值进行对比。在文本续写中，扩充词表但是没有经过继续预训练的模型的PPL值为41.287，扩充词表并经过训练以后的模型的PPL值为11.683，具体结果如下表2所示。这表明


(c) 2024 中国中文信息学会计算语言学专业委员会

继续预训练以后的模型能够有效地学习新闻文本中的语言模式并预测单词序列。下图4为扩充词表并经过继续预训练模型的续写效果展示。

| | 扩充词表未进行继续预训练模型 | 扩充词表经过继续预训练模型 |
|---|---|---|
| PPL值 | 41.287 | 11.683 |

表 2: 继续预训练前后模型的PPL值对比

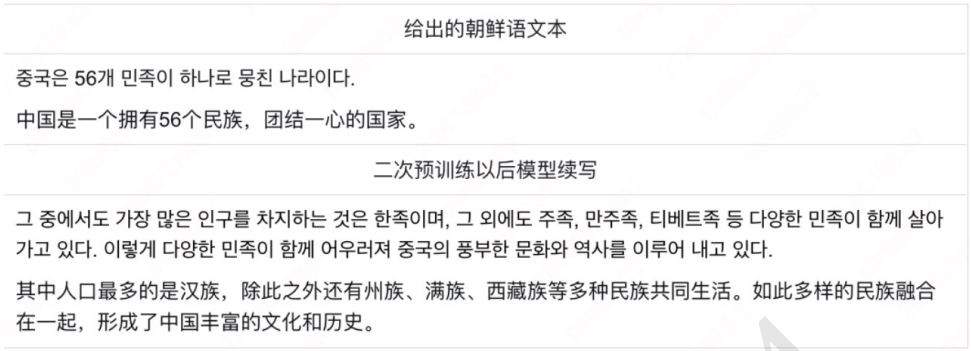



图 4 经过继续预训练后模型的续写效果

我们最终得出结论：经过预训练后的模型在文本建模任务上的良好性能表明，它可以用于各种自然语言处理应用，例如文本摘要、机器翻译和问答以及分类。进一步的研究可以探索该模型在各种数据集和任务上的性能，并对其进行微调以提高特定任务的性能。

## 4.2 指令微调实验

### 4.2.1 实验数据集

指令微调的训练数据是直接通过公开的YNAT数据按照斯坦福的Alpaca格式的Prompt模版进行拼接得到，其中训练集45678条，验证集9104条，测试集9104条，包含朝鲜语文本及其所处的类别共七个主题科技、经济、文化、美容/健康、社会、生活、世界。数据样例见图5。

| Prompt | 类别 |
|---|---|
| 아래는 작업을 설명하는 명령입니다.요청을 적절하게 완료하는 응답을 작성합니다.<br><br>###지시:<br>다음 텍스트에 대한 분류 별로 각각은 ["정치", "경제", "사회","문화","세계","IT\스포츠 과학적",""].<br>실업급여 신청은 느는데 고용은 끊겨 ...현실적인 고용 위기다<br><br>###반응:사회 | 사회 |
| 译文 | |
| 下面是描述任务的指令。编写适当完成指令的回复。<br><br>###指令:<br>对下列文本进行分类，类别分别是["政治"、"经济"、"社会''、"文化"、"世界"、"IT\科学"、"体育"]。<br>失业补贴的申请在增加，而招聘却中断....这是一场现实中的雇佣危机<br><br>###回复:社会 | 社会 |

图 5 朝鲜语文本分类Prompt样例


(c) 2024 中国中文信息学会计算语言学专业委员会

#### 4.2.2 模型评估标准

本节在Ko-LLaMA与CINO及相关模型进行分类效果评定的时候标准采用准确率(A)、精确率(P)、召回率(R)以及F1值作为实验的评价标准.计算公式为:

$$A = \frac{T_{TP} + T_{TN}}{T_{TP} + T_{TN} + T_{FP} + T_{FN}} \tag{4}$$

$$P = \frac{T_{TP}}{T_{TP} + T_{FP}} \tag{5}$$

$$R = \frac{T_{TP}}{T_{TP} + T_{FN}} \tag{6}$$

$$F1 = \frac{2 * P * R}{P + R} \tag{7}$$

式中：$T_{TP}$为样本属于类$C_i$，并被分类器正确分类到类$C_i$的样本数；$T_{FP}$为样本不属于类$C_i$，但被分类器分到类$C_i$的样本数；$T_{FN}$为样本属于类$C_i$，但被分类器分到其他类的样本数。本文所研究的朝鲜语文本分类属于多分类任务，因此采用宏平均来衡量整体的分类效果,即单独计算每个类别的P和R，再进行算术平均得到测试集总体的P和R,最后通过(7)式得到F1。

#### 4.2.3 实验设置

指令微调部分参数设置与继续预训练部分参数设置是不同的，因为在指令微调的时候，模型已经学习过朝鲜语知识，并且我们主要的任务是对齐问答格式，所以我们可以设置的相对小一些。下面的表2是指令微调时设置的参数详情。

| 模型参数 | 参数值 |
|---|---|
| Peak learning rate | $1e-4$ |
| LoRA dropout | 0.05 |
| Batch size | 1024 |
| Max sequence len | 512 |
| LoRA rank | 8 |
| Torch dtype | Float 16 |
| LoRA alpha | 32 |
| Weight decay | 0 |

表 3 指令微调训练的参数设置

本节实验分为两个阶段：

1) 验证不同数据量对经过预训练以后模型能力的影响，分别随机抽取训练集数据的1%、5%、20%、50%、80%、100%的数据对继续预训练以后的模型进行指令微调并验证效果。

2) 分别采用CINO及CINO+TextCNN、CINO+BiLSTM、CINO+TextCNN+BiLSTM（模型串联，下面称为Model series）、CINO+TextCNN/BiLSTM（TextCNN和BiLSTM并联在CINO后面，下面称为Model parallel）以及Ko-LLaMA在YNAT数据集上进行评测实验。

#### 4.2.4 实验结果分析

指令微调阶段的验证集Loss如下图6所示。和预训练阶段Loss下降非常相似，都是前期急剧下降，然后变得平缓并趋于稳定。对此做出的分析：

虽然模型已经在大量数据上预训练过，但是当模型开始在特定任务的数据上进行微调时，它仍然需要进行一些参数的调整以适应这个新任务。在微调的初始阶段，模型由于对新任务的数据分布不熟悉，因此可能会有较大的误差，这就导致了在微调初期下降得较快。而随着微调

的进行，模型对新任务的数据分布逐渐熟悉，每次更新带来的误差减小，因此Loss下降的速度会变慢。

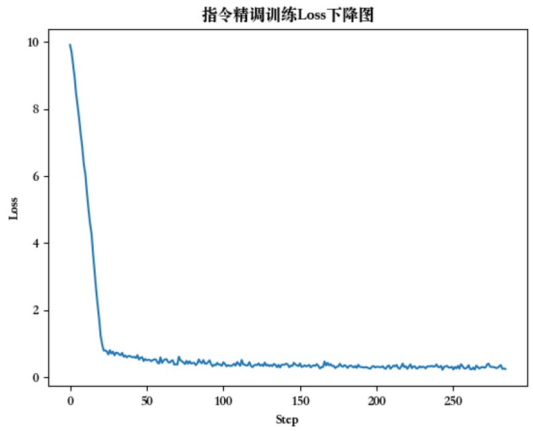



图 6 指令微调Loss损失图

第一阶段的实验：验证不同数据量对预训练以后模型能力的影响，所以我们分别随机抽取了训练数据的1%、5%、20%、50%、80%、100%进行微调，其F1-score的变化如下图7所示。

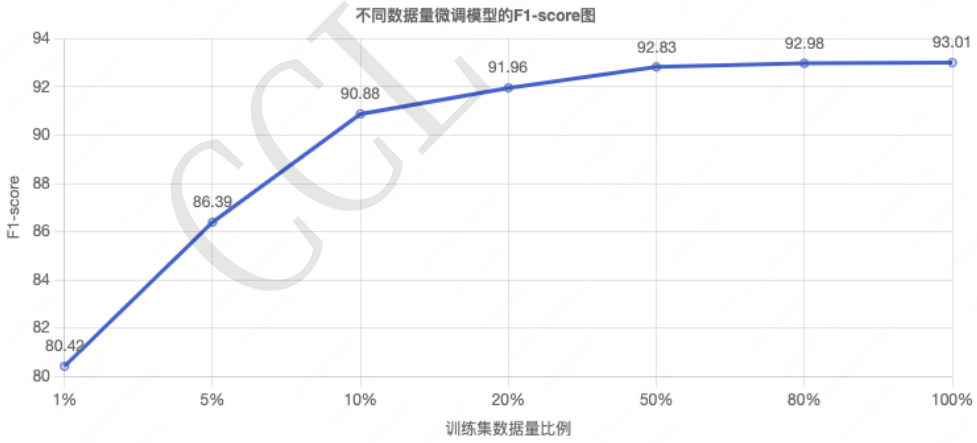



图 7 不同数据量指令微调模型的F1-score图

从图8中可以看出不同数据量对应的F1-score值的变化情况符合Loss下降的规律，当仅用1%的数据时，模型并没有很好的被充分训练，所以F1-score低，Loss高；当继续增加训练数据时，Loss急剧下降，F1-score迅速升高；再继续增加数据至50%时，F1-score就已经趋于平稳，再继续增加数据的收益将会很低。


(c) 2024 中国中文信息学会计算语言学专业委员会

表 4 不同模型分类效果的对比实验数据(单位:%)

| 模型 | 准确率(A) | 精确率(P) | 召回率(R) | F1-score |
|---|---|---|---|---|
| CINO | 87.46 | 86.57 | 88.07 | 87.29 |
| CINO+TextCNN | 86.34 | 86.42 | 86.2698 | 86.28 |
| CINO+BiLSTM | 86.62 | 86.89 | 86.77 | 86.78 |
| Model series | 87.58 | 86.15 | 88.75 | 87.43 |
| Model parallel | 87.99 | 87.58 | 89.31 | 88.44 |
| 原版LLaMA | 52.41 | 53.01 | 49.68 | 51.36 |
| GPT-3.5 | 71.33 | 72.68 | 73.59 | 73.13 |
| **Ko-LLaMA** | **92.56** | **92.83** | **93.18** | **93.01** |

第二阶段实验：使用全部指令微调数据在经过在继续预训练以后的模型上做文本分类任务的指令微微调，Ko-LLaMA在朝鲜语文本分类上的能力得到了巨大的提升。具体实验结果由表3给出，根据实验结果得出以下结论:

经过继续预训练和指令微调的Ko-LLaMA的效果要显著优于CINO的多种组合模型的效果，因为Ko-LLaMA模型参数更多并且已经在大规模的语料库上进行了预训练，当其经过词表扩充以及继续预训练，模型可以很快的学习到朝鲜语的相关知识并进行理解；同时Ko-LLaMA的效果也优于未经过词表扩充和继续预训练的大语言模型（原版LLaMA和GPT-3.5）的效果，对此的分析为原版LLaMA和GPT-3.5词表中缺少朝鲜语Token导致对朝鲜语的编码能力很弱，并且未经过朝鲜语语料的预训练更将导致对朝鲜语的理解能力不理想。最后在指令微调的数据集上，通过特定的格式的Prompt引导模型如何完成一个分类任务，这样模型可以更好的学习到如何去完成一个分类任务。最终Ko-LLaMA在YNAT朝鲜语文本分类数据集上的效果为：准确率（A）92.57%、精确率（P）92.83%、召回率（R）93.19%、F1-score 93.01%。

## 5 总结

本文提出了一种基于LLaMA并且能够理解少数民族语言-朝鲜语的大语言模型Ko-LLaMA。因为原版LLaMA无法理解和生成朝鲜语，所以我们增加了额外的2万个朝鲜语Token，并且与原版LLaMA进行了词表合并，最终词表大小为49924；然后通过爬取的58642篇朝鲜语文档进行了继续预训练，在训练过程中使用PEFT库去做带有LoRA的参数高效训练，同时使用DeepSpeed来优化训练过程中的内存效率，通过继续预训练，使Ko-LLaMA能够理解和生成朝鲜语，由于数据量的限制，朝鲜语的的理解和生成能力并不是特别强，只能进行一些基本的续写；然后我们通过公开数据集YNAT，按照斯坦福Alpaca的Prompt格式进行构造数据，并且对继续预训练以后的模型进行指令微调训练，通过随机抽取不同比例的指令微调数据进行指令微调训练，分析了不同数据量对指令微调效果的影响，经过全部数据指令微调训练以后的Ko-LLaMA在朝鲜语文本分类任务上的能力已经得到了显著提高。我们通过与CINO、CINO+TextCNN、CINO+BiLSTM、CINO+TextCNN+BiLSTM、CINO+TextCNN/BiLSTM以及原版LLaMA和GPT-3.5等一系列模型进行比较，Ko-LLaMA的朝鲜语文本分类效果比CINO等一系列模型显著提高。

## 致谢

本论文由国家社科基金重点项目(22&ZD035)，中央民族大学项目(GRSCP202316, 2023QNYL22，2024GJYY43)资助。

## 参考文献

Reza Yazdani Aminabadi, Samyam Rajbhandari, Minjia Zhang, Ammar Ahmad Awan, Cheng Li, Du Li, Elton Zheng, Jeff Rasley, Shaden Smith, Olatunji Ruwase, and Yuxiong He. 2022. Deepspeed inference: Enabling efficient inference of transformer models at unprecedented scale.


(c) 2024 中国中文信息学会计算语言学专业委员会

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

Yiming Cui, Ziqing Yang, and Xin Yao. 2024. Efficient and effective text encoding for chinese llama and alpaca.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. 2023. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Ben Mann, N Ryder, M Subbiah, J Kaplan, P Dhariwal, A Neelakantan, P Shyam, G Sastry, A Askell, S Agarwal, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

M San-Martin, M Copaira, J Zuniga, R Rodreguez, G Bustinza, and L Acosta. 1968. Aspects of reproduction in the alpaca. *Reproduction*, 16(3):395–399.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models.

CCL 2024

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions.

Mengzhou Xia, Tianyu Gao, Zhiyuan Zeng, and Danqi Chen. 2023. Sheared llama: Accelerating language model pre-training via structured pruning. *arXiv preprint arXiv:2310.06694*.

Ziqing Yang, Zihang Xu, Yiming Cui, Baoxin Wang, Min Lin, Dayong Wu, and Zhigang Chen. 2022. CINO: A Chinese minority pre-trained language model. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3937–3949, Gyeongju, Republic of Korea, October. International Committee on Computational Linguistics.

CCL 2024

# TiComR：基于提示的藏文对话型阅读理解模型

朋毛才让[1,2]　孙媛[1,2,3,*]
[1]中央民族大学　信息工程学院，北京　100081
[2]国家语言资源监测与研究少数民族语言中心
[3]民族语言智能分析与安全治理教育部重点实验室
*通讯作者：孙媛
tracy.yuan.sun@gmail.com

## 摘要

现有的对话型阅读模型在中英文对话型阅读理解任务中表现出色，但由于藏文在语法结构、表达方式等方面同中英文有显著差异，导致这些模型在对藏文对话型阅读理解的对话历史进行建模时存在困难。鉴于此，本文利用当前大模型的优越能力，提出了一种基于提示的对话历史建模方法–TicomR，以解决藏文对话型阅读理解任务中模型性能受限的问题。该方法通过引入基于提示的学习机制，直接在段落文本中添加提示来突显对话历史，而非修改段落标记嵌入，从而在微调过程中实现对对话历史的精确建模，以增强模型对问题的理解能力。实验结果表明，TiComR模型在藏文对话型阅读理解任务上取得了显著的性能提升，并在英文数据集CoQA上也有较好的表现。本文将TicomR开放供研究使用，https://github.com/Tshor/TicomR。

**关键词：** 藏文；对话型；文本理解；提示学习；大模型

# TiComR: A Prompt-based Tibetan Conversational Reading Comprehension Model

**PengmaoCairang[1,2] Yuan Sun[1,2,3,*]**
[1] School of Information Engineering, Minzu University of China, Beijing　100081
[2] National Language Resources Monitoring and Research Center for Minority Languages
[3] Key Laboratory of Ethnic Language Intelligent Analysis and Security Governance of MOE
*Corresponding author: Yuan Sun
tracy.yuan.sun@gmail.com

## Abstract

Existing conversational reading models excel in Chinese and English conversational reading comprehension tasks. However, these models struggle with Tibetan conversational reading comprehension due to significant differences in grammatical structure and expression between Tibetan and Chinese/English. To address this issue, this paper introduces TicomR, a prompt-based conversational history modeling method that leverages the advanced capabilities of current large models. Instead of modifying paragraph token embeddings, TicomR enhances the modeling of conversational history by directly integrating prompts within the paragraph text, allowing for precise fine-tuning and improved understanding of questions.Experimental results demonstrate that the TiComR model significantly improves performance in Tibetan conversational reading comprehension tasks and also performs well on the English CoQA dataset. TicomR is available for research use at https://github.com/Tshor/TicomR.

**Keywords:** Tibetan , Conversational , Text comprehension , Prompt learning , Large model

第二十三届中国计算语言学大会论文集，第242页–第253页，太原，中国，2024年7月25日至28日。
卷1：主会论文
(c) 2024 中国中文信息学会计算语言学专业委员会

## 1 引言

近年来，由于Siri、Alexa和Google Assistant等语音助手平台的广泛应用以及深度学习的进展，对话式问答（Conversational Question Answering, ConvQA）引起了广泛关注 (Rashid et al., 2024)。给定一个文本段落和一段对话，ConvQA的目标是从段落中提取出对话的最后一个问题的答案。目前，大规模的ConvQA基准已经提出，如CoQA (Reddy et al., 2019)、QuAC (Choi et al., 2018)。从技术上讲，ConvQA作为问答（Question Answering, QA）任务的扩展，与传统的单轮问答任务相比，其显著特点是它包含了对话历史，这对于有效建模历史信息提出了挑战。过往研究表明，简单地将对话轮次拼接到输入序列中的方法存在局限 (Gupta et al., 2020)。因此，学者们提出了多种显式建模对话历史的架构组件方案，如HAE (Qu et al., 2019a)、HisBERT (Liu et al., 2020)等。多数研究仅基于特定的CQA基准数据集，如CoQA或QuAC汇报了核心指标。尽管研究人员已在各项基准数据集上取得了显著的成功，但仍然存在许多在以前的ConvQA努力中很少提到的重要障碍：现有的ConvQA数据集主要建立中英等主流高资源语言上，不支持藏文等低资源语言在对话式问答领域的探索；以往的ConvQA模型仅建立在英语语料库上，限制了其他语言ConvQA的发展，如Xlnet (Yang et al., 2019)、SDNet (Zhu et al., 2018)等；藏语独特的语法结构、习惯表达和文化背景，使得主流模型难以迁移，甚至对于LLMs来说也很难完全捕捉到，这通常会导致误解或生成听起来不自然的文本。

以ChatGPT (Ouyang et al., 2022)、Llama (Touvron et al., 2023)等为代表的大语言模型(Large Language Models, LLMs)展现出的理解和生成类人文本能力，显著推动了自然语言处理（Natural Language Processing, NLP）领域的进步与发展 (Moradbeiki and Ghadiri, 2024)。通过最新的研究发现，各种任务之间可能存在紧密的联系，这揭示了一种统一的范式，可以潜在地应用于解决各种QA任务，以模拟它们的共性 (Zhong et al., 2022)。然而，由于大规模藏语训练数据的有限可用性和语言的固有特征，将LLMs直接应用于藏语ConvQA任务面临障碍。

为了应对这些挑战，本文跟随Gekhman等人 (Gekhman et al., 2023)的思想提出了一种依靠大模型能力的基于提示学习的藏文对话式阅读理解模型–TicomR（Tibetan Conversational Comprehension Reader），利用大语言模型的知识和能力来作为基础，微调时在基础文档中添加文本提示，以便突出显示对话历史中以前的答案。实验结果表明，TicomR在TiconvQA (DawaCairen et al., 2024)上表现优异，F1值达到82.19，相比最好的基准有15.69个F1值的相对提升，并在CoQA上也有不错的表现。

本文的主要贡献如下:

(1)本文提出一种利用大模型能力作为基础，并采用提示学习方法进行构建的藏文对话式阅读理解模型–TicomR。通过结合大模型的优势和提示学习的灵活性，TicomR模型能够更有效地捕捉藏文对话的复杂性和特殊性，从而提升对话式阅读理解任务的准确性。

(2)TicomR可以利用少量的数据资源，结合高效低成本的提示学习方法显著提升模型性能，这为藏语等低资源语言的对话型问答任务提供了一个有效的解决方案。

(3)实验结果表明，TicomR在TiconvQA以及CoQA数据集上均表现优异。在TiconvQA上模型最好实验结果获得了82.19%的F1值，相比基线模型提升了15.69%。在CoQA上，得到78.12%的F1，相比基线模型提升了1.52个F1值。此外，本文已将TicomR开源至：https://github.com/Tshor/TicomR。

## 2 相关工作

机器阅读理解（Machine Reading Comprehension, MRC）任务是NLP领域的一个重要分支，它旨在通过理解文本内容来自动回答关于该文本的问题。MRC任务的核心在于系统能否准确地从给定的文本中检索或生成答案，以响应特定的问题 (Rashid et al., 2024)。这一任务最初由文本检索会议（Text REtrieval Conference, TREC）提出 (Voorhees et al., 1999)，目的是通过查询检索相应的文本段落来解答问题。随着研究的深入和技术的进步，MRC任务已经从最初的简单检索扩展到了更加复杂的对话式问答。与传统的单轮MRC不同，ConvQA涉及多轮对话，系统不仅需要理解当前的问题和上下文，还需要记住之前的对话历史，以便更准确地生成或检索答案 (Rashid et al., 2024)。这种多轮对话的模式更接近人类的交流方式，因此对系统的

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

理解和记忆能力提出了更高的要求。为了应对ConvQA的挑战，研究者开发了如CoQA (Reddy et al., 2019)，DoQA (Campos et al., 2020)等数据集，这些数据集要求模型理解和利用对话历史来更准确地回答问题。

对话历史建模是对话式问答中的一个主要挑战。早期的解决方案主要依赖于递归神经网络（RNN）和注意力机制的变体来处理序列数据和捕捉长距离依赖关系 (Reddy et al., 2019; Choi et al., 2018; Zhu et al., 2018);另一个趋势是使用基于流的方法，使用对话历史的向量表示在段落中产生潜在的表示 (Huang et al., 2018; Yeh and Chen, 2019; Chen et al., 2021)。然而，随着深度学习的发展，基于Transformer模型的预训练语言模型，如BERT (Devlin et al., 2019)和GPT (Radford et al., )，已经成为了处理这类任务的主流方法。这些模型通过在大规模语料库上进行预训练，学习到了丰富的语言表示，能够更好地理解和生成自然语言。为了更好地对对话历史进行建模，研究者们尝试了多种方法。例如，FlowQA (Huang et al., 2018)和GraphFlow (Chen et al., 2021)使用每个单词作为图形中的节点，并使用注意机制来表示历史；HAE (Qu et al., 2019a)认为历史基础答案是上下文，这对于现实生活中的对话代理来说是不切实际的；Pos-HAE (Qu et al., 2019b)将历史转向位置视为附加编码；还有基于回溯 (Qiu et al., 2021)和基于查询重写 (Vakulenko et al., 2021)等方法的模型。这些方法通常涉及到对历史信息的编码和更新，以便在生成答案时能够考虑到整个对话的上下文。这种全局视角对于理解和生成连贯、一致的多轮对话至关重要。

在过去的几年中，大型语言模型在NLP的多个领域取得了显著的成就。这些模型依赖于深层的神经网络架构和大规模的预训练数据集，从而在多种NLP任务中展现出卓越的性能。LLMs之所以能够取得这些成果，很大程度上归功于它们在预训练阶段学习到的强大语言表示能力 (Zhao et al., 2023)。LLMs在各种NLP任务中的应用非常广泛，包括但不限于文本生成、情感分析、实体识别和机器翻译等。这些任务通常首先基于预训练的模型进行构建，随后通过微调过程进一步优化模型参数，以更好地适应特定的任务需求。微调过程中，模型在特定任务的数据集上进行训练，从而调整预训练参数以提高任务性能。近期的研究趋势表明，面向提示的微调方法正逐渐受到研究者的关注 (Liu et al., 2023)。这种方法通过在输入文本中插入硬提示标记，如手动设计的标记 (Schick and Schütze, 2021)或自动搜索得到的标记 (Jiang et al., 2020)，来引导模型关注输入文本的关键部分。这些硬提示标记类似于预训练任务中的提示，但它们被重新设计以适应下游任务的目标。然而，考虑到硬提示标记在连续的嵌入空间中可能不是最优解，最新的研究开始探索可调软提示的概念 (Zhong et al., 2022)。软提示通过引入可学习的参数来调整输入文本的表示，从而在模型尺寸较大时实现更加令人满意的性能 (Lester et al., 2021)。与大型预训练语言模型（Pretrain Language Models, PLMs）中的复杂参数相比，软提示提供了一种轻量级且可插拔的解决方案，这为适应新的问答任务提供了更大的灵活性。

与英语等被广泛研究的语言相比，对藏语对话型阅读理解的研究仍处于早期阶段。现有的研究主要集中在单轮抽取式机器阅读理解任务 (Sun et al., 2021)，受限于低资源数据集的稀少、藏语语法规则、文化背景差异等原因，主流模型无法直接迁移至藏文。当前模型在处理高度语境化的问题时，面临着自然语言固有的模糊性挑战，尤其是需要深入挖掘隐含信息，超越表层问题以更深入理解上下文 (Moradbeiki and Ghadiri, 2024)。这些局限性凸显了对研究更加适应或更加稳健的藏语ConvQA模型的迫切需求。

## 3 TicomR模型

受到当前大语言模型在各项任务上展现出来的优异性能，以及提示学习方法的高效性启发，本文提出了TicomR。这是一种基于提示的对话历史建模方法，该方法通过在段落中的相应位置插入文本提示来突出显示先前对话回合中答案的证据文本，而不是嵌入对话历史的向量表示来对其编码。我们期望能够引导大模型更加聚焦于有价值的对话内容，同时降低现有基于嵌入方法所带来的学习复杂性，进而提升模型在对话处理任务中的性能与效率。

### 3.1 任务定义

给定一个文本段落$P$，当前问题$Q_k$ 以及以一系列先前问题和答案的形式存在的对话历史$H_k = (Q_1, A_1, Q_2, A_2, \ldots, Q_{k-1}, A_{k-1})$，一个对话式问答（ConvQA）模型基于文本段落$P$ 和对话历史$H_k$作为知识源来预测答案$A_k$。这些答案可以是段落$P$ 内的文本片段（抽取式）或自由格式的文本（生成式）。


(c) 2024 中国中文信息学会计算语言学专业委员会

### 3.2 面向藏文对话型阅读理解的提示方法

TicomR采纳了标准的多轮问答模型框架和输入机制，其输入涵盖了当前问题$Q_k$和段落$P$以及对话历史$H_k$。针对每个CQA示例$(P, H_k, Q_k)$，TicomR会基于先前的对话内容，在段落$P$中插入文本提示。在抽取式问答的情境下，答案$A_k$往往是段落$P$中的某个片段。当给定输入$(P, H_k, Q_k)$时，TicomR会将段落$P$转化为带有证据文本高亮显示的版本$P^*$（通过创建提示并嵌入到$P$中）。这些提示由定位$P$中所有历史答案的起始和终止位置组成，并在相应位置插入独特的文本标识符（详见图1中的实例）。随后，经过处理的输入$P^*$将传递给大模型进行微调，而非原始段落。在生成式问答的场景中，TicomR首先从段落$P$中提取证据范围，随后生成自由文本形式的答案。因此，最终答案可能并不直接出现在$P$中。为支持这一情境，TicomR会突出显示历史证据范围（存在于$P$中的部分），而非直接生成的答案。

为编码对话中的位置信息，答案的标记采用了其逆序转置索引号，即$k-1-j$。这种编码方式反映了历史答案相对于当前问题$Q_k$的位置，使得模型能够依据它们的顺序来区分不同的历史答案(例如图1中的<1>标引之间的文本即为相对于当前问题Q9的第一轮对话历史，也就是第8轮对话的证据文本)。在CQA对话中，有时会遇到无法回答的问题。对于包含这类问题的文本，TicomR在插入提示之前，会先在$P$中附加一个“NO ANSWER”的字符串。随后，与常规历史答案相似，每个历史“NO ANSWER”也会被突出显示。例如，请参见图1中的$A_9$。

总体而言，TicomR相较于以往的提示方法具备以下特征：（1）标记化形式，使用尖括号来标记提示或指示，这种标记化形式使得对话历史更加清晰可辨（例如，“<1>”）。尽管大多数文本提示方法利用由自然语言组成的提示，但我们的提示包含非言语符号，这些符号已被证明对NLP任务的监督具有用处。例如，Aghajanyan等人 (Aghajanyan et al., 2021)通过向输入文本中添加HTML符号展示了结构化预训练的实用性。（2）数字编码，每个标记中的数字同对话历史的轮数关联，引导模型找出当前问题对应的证据文本，以使模型具备当检索定位到多个相同的证据文本时，明确当前问题的证据文本位置。（3）动态性，与大多数预定义提示在输入中位置的提示方法相比 (Liu et al., 2023)，我们的提示会针对每个示例插入到不同的位置。（4）缺失处理，引入<NO ANSWER>标记用于处理证据文本缺失的情况，使得整体提示更具可读性和完整性。这是由于TiconvQA中包含抽取式、是否类以及不可回答类问题，这使得答案和证据文本有可能并不是直接存在于文本中。

### 3.3 提示设计

如上所述，TicomR在段落中每个历史答案的开始和结束位置插入提示（图1）。提示设计有预定义的标记符号，并包括答案的回合索引（例如，“<1>”）。该设计主要考虑到以下两点：首先，文本提示可以表示对话历史信息；其次，提示在P中的定位有助于引导模型找出当前问题对应的证据文本，以使模型具备当检索定位到多个相同的证据文本时，明确当前问题的证据文本位置。

我们将输入修改为段落和当前问题，而不保留对话历史。这使得模型只能依赖在文本中证据标引来获取历史信息，实验结果表明，TicomR确实编码了来自对话历史的信息，因为显性对话历史的输入与否对于模型性能有显著变化。另外，我们将证据文本在文本中的提示调整为“<>”，即没有转向索引，我们发现，这使得模型性能下降了3.76%，这表明在段落中的提示位置至关重要，并且TicomR的大部分性能增益来自于其提示相对于历史答案的位置。此外，我们将提示策略修改为插入相同数量的符号，但在段落中的相同证据文本但不同位置或随机位置。我们发现模型性能下降了5.08%，表明只有当提示被插入到有意义的位置时，模型才能学会利用这些位置以得到有效的历史表示。

最后，由于文本提示允许轻松注入额外信息，Gekhman等人 (Gekhman et al., 2023)在这个方向上进行了几次初始尝试，将不同类型的信息注入到文本提示中，它们的实验结果表明，提示应该保持简洁符号并在段落中占用小空间的证据。本文尝试使用藏文数字来进行标引，查看同文本一致的提示类型是否有助于提升模型性能，实验结果表明引起的变化并不显著。

### 3.4 基于大模型的对话型阅读理解微调

大模型微调方法是一种针对预训练模型的技术，主要用于提升预训练模型在特定任务或领域上的性能。本文选择TiLamb (Zhuang et al., 2024)作为基座模型进行藏文对话型阅读理解任务微调及一系列工作。这是一个基于LLaMA2-7B增量预训练的藏文大语言模


(c) 2024 中国中文信息学会计算语言学专业委员会

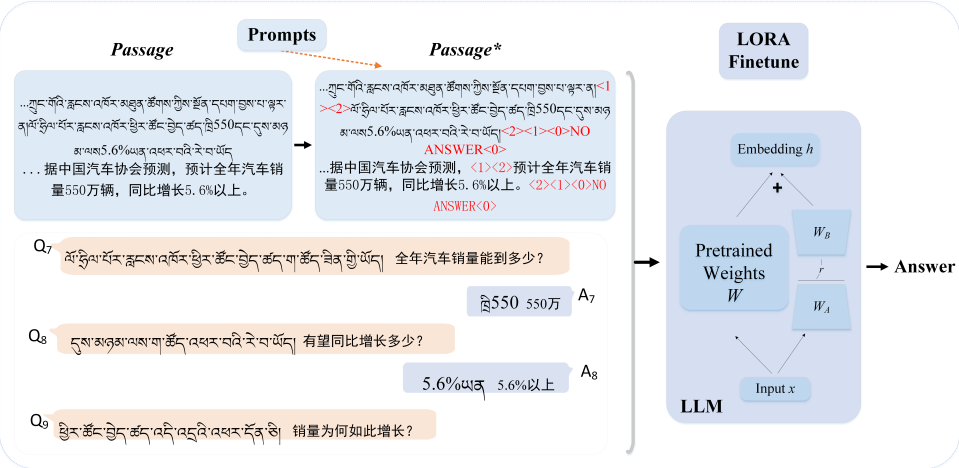



图 1: TicomR中提示突出显示方法

型，它使用了26.43GB的藏文语料进行了增量预训练，并在LLaMA2的基础上扩展了词表，从原有的32000扩充藏文词汇至61221。具体而言，扩展词汇表的过程包括以下步骤：首先，我们使用SentencePiece工具的BPE算法在藏文语料库上训练了一个词表大小为32,000的藏文分词模型；其次，将藏文分词模型与LLaMA原生的分词模型进行合并，通过将藏文tokenizer的tokens添加到LLaMA tokenizer中，创建一个新的tokenizer，该tokenizer能够同时处理英文和藏文文本；最后，排除重复的token后，对词表新增的token进行初始化，并采用均值扩充的方法赋以初值。此外，TiLamb还对LLaMA2-7B原始模型的embedding和lm_head进行了均值扩充初始化。实验结果表明，TiLamb在多个藏文NLP下游任务上展示了出色的性能，例如藏文新闻分类、藏文实体关系分类、藏文机器阅读理解等。TiLamb不仅继承了Llama2的强大能力，而且针对藏文进行了特定的优化和增量预训练，这使得我们的方法能够更有效地应用于藏文对话型阅读理解任务。这一选择是出于两个方面的考虑：首先，尽管LLaMA、Qwen等大模型在中英文等主流语言上表现出色，但并不支持藏文等低资源语言，这限制了我们在藏文对话型阅读理解任务上直接利用主流LLM的能力。其次，TiLamb作为专门为藏文设计的增量预训练大模型，能够更好地捕捉藏文的语法结构和表达方式，从而更准确地建模藏文对话历史。

本文选择LoRA（Low-Rank Adaptation）方法 (Hu et al., 2021)进行TiLamb的藏文对话型阅读理解任务微调。LoRA作为一种高效的大模型微调方法，其核心思想是通过在其权重矩阵中引入低秩矩阵来适应预训练的语言模型。该低秩自适应层初始化随机值，并在微调过程中更新。具体来说，LoRA首先从一个经过预训练的语言模型开始，在其权重矩阵中添加一个低秩适应层。这一低秩适应层以随机初始化的低秩矩阵形式呈现，为模型提供了灵活性和适应性。随后，在新任务或领域的训练过程中仅专注于更新这一低秩适应层，而预训练模型的原始权重则保持不变。这允许模型在不改变其一般知识的情况下有效地学习特定于任务的信息。最终，然后利用经过微调的模型，使用经过调整的模型对来自目标任务或领域的新数据进行预测。通过将适应过程集中在低秩矩阵上，LoRA显著提升了微调的效率，相较于全面微调，大幅减少了计算和内存的开销。在数学上，假设输入是$x$，输出是$h$，预训练模型的权重是$W_0$，降维矩阵是$A$，升维矩阵是$B$。采用LoRA微调时，公式如（1）所示：

$$h = W_0 x + \Delta W x = W_0 x + BAx \tag{1}$$

其中，$\Delta W$是需要更新的参数，它的分解意味着我们需要用两个较小的LoRA矩阵A和B来表示较大的矩阵。


（c）2024 中国中文信息学会计算语言学专业委员会

# 4 实验评估

## 4.1 数据集

本文在TiconvQA (DawaCairen et al., 2024)和CoQA (Reddy et al., 2019)上进行了实验，将数据集按照8：2的比例随机划分为训练集、测试集，表1是这些数据集的统计信息。

| dataset | passage | Q/A pairs | passage length | turns per passage |
|---|---|---|---|---|
| TiconvQA | 2120 | 20358 | 198 | 9.6 |
| CoQA | 8399 | 127k | 271 | 15.2 |

表 1: 实验数据集信息

TiconvQA：该数据集是一个藏文对话型阅读理解数据集，通过人工和半自动方法进行构建。TiconvQA拥有来自2120篇地理、人物和新闻三个领域的文本产生的20138轮对话。其中地理和人物领域的文本来自云藏百科，新闻文本来自人民网藏版等权威新闻网站。

CoQA:CoQA数据集收集了来自七个不同领域的8,000个对话，共计126,000个问题，主要来源于CNN等新闻文章。对话构建过程中，采用了两名注释者，一个提问一个回答，都涉及整个上下文。问题是自由陈述，但需要足够的证据和可用的推理。答案是自由形式的，并在文章中强调了相应的基本原理。

## 4.2 评估指标

为了评估模型的效果，本文使用EM值和F1值两个指标进行评价。

EM值：EM值（Exact Match）是一个关键指标，用于衡量模型预测的答案与标准答案之间的完全匹配程度。它反映了模型在精确捕获问题相关细节和语义方面的能力，是评估模型性能的重要标准之一。高EM值通常意味着模型在理解和回答问题方面具有更高的准确性，具体计算方法如公式（2）。

$$EM = \frac{\text{完全匹配的样本数}}{\text{总样本数}} \tag{2}$$

其中，完全匹配的样本数是指模型输出与参考输出完全一致的样本数量。总样本数是指评估过程中使用的样本总数。

F1值：在对话式阅读理解中，F1值作为评估指标，通过计算精确率和召回率的调和平均数，综合衡量模型性能。它既能体现模型预测的准确性，又能反映模型对正例的覆盖能力，是评估对话式阅读理解模型效果的关键指标，如公式（3）-（5）所示。

*Precision*（精确度）衡量预测为正例的样本中真正为正例的比例，计算公式为：

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

*Recall*（召回率）衡量所有真正正例中被模型预测出来的比例，计算公式为：

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

*F1*则是Precision和Recall的调和平均值，用于综合评估模型的性能，计算公式为：

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{5}$$

其中，真正例（True Positive, TP）表示实际为正例且预测也为正例的样本数，假正例（False Positive, FP）表示实际为反例但预测为正例的样本数，假反例（False Negative, FN）表示实际为正例但预测为反例的样本数。


(c) 2024 中国中文信息学会计算语言学专业委员会

| Parameters | Values |
|---|---|
| cutoff_len | 2048 |
| learning_rate | 2e-4 |
| num_tran_epochs | 2.0 |
| max_samples | 100000 |
| per_devices_train_batch_size | 4 |
| gradient_accumulation_steps | 4 |
| max_grad_norm | 1.0 |
| lora_rank | 8 |
| lora_dropout | 0.05 |
| resume_lora_training | True |

表 2: TicomR模型实验参数具体细节

### 4.3 实验参数设置

我们对选取的基线模型的实验参数均遵循原始论文的最好设置，在此不做详细描述。对于TicomR的具体实验参数设置如表2所示，所有实验均在2张Tesla V100-PCIE-32G上完成。在表3中我们比较了基线模型在TiconvQA和CoQA上的结果。

### 4.4 在经典数据集上的表现

本文选择DrQA (Chen et al., 2017)、SDNet (Zhu et al., 2018)、TiBERT (Liu et al., 2022)、TBERT (Dslab-NLP, 2023)等作为基线模型，以上前两者是英文领域经典模型，TiBERT、TBERT是现有常用的藏文预训练语言模型，它们在英文、藏文领域文本分类、情感分析等下游任务上有着出色的表现。具体实验结果表3所示。

（1）DrQA：DrQA是一种基于深度学习的问答模型，通过双向长短期记忆网络（BiLSTM）编码文档和问题，并结合注意力机制。其特点包括端到端的训练、支持多种问答任务、良好的扩展性和高效性，以及通过交互式学习进一步提高性能。

（2）SDNet：SDNet是一种用于端到端问答的神经模型，它通过联合学习问题理解、证据检索和答案生成的过程来提高性能。模型采用了一种层次化的注意力机制，首先对问题进行编码，然后从支持文档中检索相关信息，最后利用这些信息生成准确的答案。SDNet在多个问答数据集上进行了评估，显示出优于传统方法和现有神经模型的性能。

（3）TiBERT：TiBERT是针对藏语自然语言处理任务设计的预训练语言模型。它在大规模藏语训练数据上进行训练，利用Sentencepiece构建了能够覆盖大部分藏语单词的词汇库。TiBERT在文本分类和问题生成等下游任务上展现了出色的性能，并相比经典模型和多语言预训练模型具有优势。该模型的推出为藏语自然语言处理领域的发展提供了有力支持。

（4）TBERT：TBERT（Tibetan-BERT）模型是一种针对藏文自然语言处理任务设计的预训练语言模型。由青海师范大学省部共建藏语智能信息处理及应用国家重点实验室的多拉教授团队与兰州大学开源软件与实时系统教育部工程研究中心共同开发。TBERT旨在解决藏语在自然语言处理领域的数据资源限制和技术挑战，推动藏文信息处理技术的发展。模型基于BERT架构，使用SentencePiece分词器，适用于各种藏语NLP任务。

TicomR在CoQA和TiconvQA两个数据集上均展现出良好的性能。在CoQA数据集上，TicomR的EM和F1值分别达到了50.6%和78.1%。与SDNet相比，TicomR在在F1值上提升了1.5%；与DrQA相比，TicomR在F1值上提升了22.5%。而在TiconvQA数据集上，TicomR的表现亦优于其他模型，其EM值达到了58.9%，F1值达到了82.1%。具体而言，与SDNet相比，TicomR在F1值上提升了35.9%；与DrQA相比，TicomR在F1值上提升了15.6%。

### 4.5 不同提示方法对模型性能的影响

为了深入探讨不同提示方法对模型性能的影响，我们在实验中尝试了是否添加对话历史($H_k$)、对话历史位置空缺提示(<>)、随机位置标引提示(random<>)、藏文数字标引提示(<Tibetan number>)以及对话历史位置相对当前问题逆向标引提示(TicomR)等多种策略，


(c) 2024 中国中文信息学会计算语言学专业委员会

| Model | TiconvQA | | CoQA | |
|---|---|---|---|---|
| | F1 (%) | EM (%) | F1 (%) | EM (%) |
| 人类表现 | 89.5 | 80.2 | 88.8 | - |
| DrQA | 66.5 | 44.5 | 55.6 | 46.2 |
| SDNet | 46.2 | - | 76.6 | - |
| TiBERT | 40.6 | 32.7 | - | - |
| TBERT | 37.4 | 21.8 | - | - |
| TicomR | **82.1** | **58.9** | 78.1 | 50.6 |

表 3: TicomR在不同数据集上的表现

并对比了它们在TiComR模型上的表现。实验结果表明，逆向标引提示方法在TiComR模型上取得了最佳性能。如表4所示，相较于其他提示方法，逆向标引提示在F1和EM指标上均表现出了明显的优势。这种提示方法充分利用了对话的连贯性和上下文信息，有效提升了模型对文本的理解和响应能力。

| Prompts | F1 | EM |
|---|---|---|
| $H_k$ | 80.39 | 56.82 |
| <> | 78.43 | 52.26 |
| random<> | 77.11 | 54.48 |
| <Tibetan number> | 81.12 | 57.40 |
| TicomR | **82.19** | **58.92** |

表 4: 不同提示对于模型性能影响

具体而言，尽管对话历史已通过提示方法标引在文章里，但仍然提供原始对话历史有助于模型更好地理解对话轮之间的潜在联系。而只保留<>符号不提供相对位置使得模型无法清晰地定位对应对话历史，虽然简单直接，但可能无法充分捕捉对话的语义信息，导致模型性能有显著下降，这也证明了对话历史的相对位置对于模型理解对话轮之间的关联非常重要。随机文本提示则由于其随机性，无法为模型提供稳定且有效的上下文信息，并有可能导致模型接收到无意义的噪声信息，进一步影响性能。随后，我们尝试了使用藏文数字标引提示来指示对话历史的位置。这种方法相较于前两种提示方法有所改进，但仍然不是最优解。尽管藏文数字为模型提供了位置信息，但可能由于模型对于藏文数字标引并不敏感，只是将其理解为文本信息。相比之下，逆向标引提示能够根据对话的实际情况动态生成提示信息，从而更好地引导模型理解和生成符合语境的响应。这种方法不仅提高了模型的准确率，还增强了模型的鲁棒性，使其能够更好地应对各种复杂的对话场景。

综上所述，不同提示方法对模型性能具有显著影响。在选择提示方法时，我们需要综合考虑对话的连贯性、语义结构以及历史信息的利用方式。通过设计合理的上下文相关提示，我们可以有效提升对话型阅读理解模型的性能。

### 4.6 消融实验

在本节中，我们进行了一系列消融实验，以验证TicomR模型中各个组件的有效性。实验结果表明，在引入提示后，TicomR模型的性能得到了显著提升，这证明了提示机制在增强模型性能方面的重要作用。具体而言，如表5所示，引入提示后，模型在EM和F1指标上均获得了显著提升。与仅使用微调（Finetune）的模型相比，添加提示（Prompts）的模型在EM和F1指标上分别提高了3.8和5.05个百分点。这表明提示能够有效地帮助模型捕捉和理解对话中的关键信息，从而提高了回答的准确性。

进一步地，当我们综合考虑TicomR模型的所有组件时，其性能得到了进一步的提升。与仅添加提示的模型相比，完整的TicomR模型在EM和F1指标上分别增加了5.68和6.26个百分点。这一结果表明，TicomR模型中的各个组件相互协作，它们共同增强了模型对对话语境的捕捉和


(c) 2024 中国中文信息学会计算语言学专业委员会

| Model | EM | Increase | F1 | Increase |
|---|---|---|---|---|
| +Finetune | 53.24 | - | 75.91 | - |
| +Prompts | 57.04 | +3.8 | 80.96 | +5.05 |
| TicomR | 58.92 | +5.68 | 82.19 | +6.26 |

表 5: TicomR的消融实验

理解能力。

通过消融实验，我们深入了解了TicomR模型中各个组件的贡献。这些结果不仅验证了提示机制的有效性，还为我们提供了关于如何进一步优化模型的宝贵启示。在未来的工作中，我们将继续探索更多有效的提示策略和方法，以进一步提升TicomR模型的性能。此外，值得注意的是，尽管提示机制在增强模型性能方面发挥了重要作用，但可能并不是唯一的因素。在实际应用中，我们还需要综合考虑如模型的架构、训练数据的质量和数量等，以全面提升模型的性能。因此，在未来的研究中，我们将继续探索这些因素对模型性能的影响，并寻求更优化的解决方案。

### 4.7 不同提示策略下历史长度的影响

在本节中，我们探讨了不同长度的对话历史信息对模型性能的影响。我们将不同长度的对话历史信息与当前问题进行拼接，分别考虑了将当前问题前h轮的对话历史信息拼接到输入中的情况，其中h的取值包括h=0、h=2、h=4以及h=all四种情况。这里的h=all表示将当前问题之前的所有对话历史信息都进行拼接。我们使用了两种不同的提示策略：正向提示和逆向提示，并将它们与原始输入进行了对比。实验结果如表6所示。从表中数据可以看出，无论是采用正向提示和逆向提示，模型在引入对话历史信息后性能均有所提升，这验证了对话历史信息在对话问答中的有效性。

| Strategy | | h0 | | h2 | | h4 | | h-all | |
|---|---|---|---|---|---|---|---|---|---|
| | | F1 | EM | F1 | EM | F1 | EM | F1 | EM |
| original | | 73.32 | 51.17 | 75.91 | 53.24 | 76.08 | 53.46 | 75.87 | 53.01 |
| prompt | reverse | 79.51 | 56.52 | 80.96 | 57.84 | **82.19** | **58.92** | 81.95 | 56.55 |
| | forward | 79.42 | 55.89 | 80.69 | 57.18 | 81.46 | 58.12 | 80.43 | 56.21 |

表 6: 不同提示策略下的实验结果

进一步地，我们观察到对于逆向和正向两种提示策略，它们在利用对话历史信息方面表现出不同的特点。具体而言，逆向提示策略在多数情况下优于正向提示策略。这可能是因为逆向提示更能有效地捕捉对话历史中的关键信息，因为它按照从最近到最远的顺序标引对话历史，使得模型更能够关注到与当前问题最为相关的历史信息。另外，我们观察到当h取值为4时，模型性能达到最佳。这表明适当数量的历史信息对于模型性能至关重要。过多的历史信息（如h=all）可能导致模型在处理时引入噪声，而过少的历史信息则可能无法提供足够的上下文信息。因此，选择合适的h值对于平衡信息量和模型性能至关重要。

## 5 总结与展望

本文针对藏文对话型阅读理解任务中模型性能不足的问题，提出了一种新型的基于提示的藏文对话型阅读理解模型——TiComR。通过结合当前大模型的优异能力和基于提示的学习机制，TiComR显著提升了在藏文对话型阅读理解任务上的性能。实验结果表明，TiComR模型在TiconvQA和CoQA数据集上都取得了优异的表现，未来可以进一步扩充训练数据集、优化模型结构，并将TiComR模型应用到其他低资源语言的对话型阅读理解任务中，以验证其通用性和适用性。


(c) 2024 中国中文信息学会计算语言学专业委员会

## 致谢

本论文得到了国家社科基金(22ZD035), 国家自然科学基金(61972436), 中央民族大学项目(GRSCP202316, 2023QNYL22, 2024GJYY43)资助。

## 参考文献

Armen Aghajanyan, Dmytro Okhonko, Mike Lewis, Mandar Joshi, Hu Xu, Gargi Ghosh, and Luke Zettlemoyer. 2021. Htlm: Hyper-text pre-training and prompting of language models. In *International Conference on Learning Representations*.

Jon Ander Campos, Arantxa Otegi, Aitor Soroa, Jan Deriu, Mark Cieliebak, and Eneko Agirre. 2020. Doqa–accessing domain-specific faqs via conversational qa. *arXiv preprint arXiv:2005.01328*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879.

Yu Chen, Lingfei Wu, and Mohammed J Zaki. 2021. Graphflow: exploiting conversation flow with graph neural networks for conversational machine comprehension. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 1230–1236.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184.

DawaCairen, PengmaoCairang, and Yuan Sun. 2024. 面向对话式阅读理解的高质量藏语数据集构建(construction of high-quality tibetan language dataset for conversational reading comprehension). In *Proceedings of the 23th Chinese National Conference on Computational Linguistics*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Dslab-NLP. 2023. Tibetan-plm. https://github.com/Dslab-NLP/Tibetan-PLM.

Zorik Gekhman, Nadav Oved, Orgad Keller, Idan Szpektor, and Roi Reichart. 2023. On the robustness of dialogue history representation in conversational question answering: A comprehensive study and a new prompt-based method. *Transactions of the Association for Computational Linguistics*, 11:351–366.

Somil Gupta, Bhanu Pratap Singh Rawat, and Hong Yu. 2020. Conversational machine comprehension: a literature review. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2739–2753.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Hsin-Yuan Huang, Eunsol Choi, and Wen-tau Yih. 2018. Flowqa: Grasping flow in history for conversational machine comprehension. In *International Conference on Learning Representations*.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059.

Chuang Liu, Deyi Xiong, Yuxiang Jia, Hongying Zan, and Changjian Hu. 2020. Hisbert for conversational reading comprehension. In *2020 International Conference on Asian Language Processing (IALP)*, pages 147–152. IEEE.


(c) 2024 中国中文信息学会计算语言学专业委员会

Sisi Liu, Junjie Deng, Yuan Sun, and Xiaobing Zhao. 2022. Tibert: Tibetan pre-trained language model. In *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2956–2961. IEEE.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Pardis Moradbeiki and Nasser Ghadiri. 2024. Perkwe_coqa: enhance persian conversational question answering by combining contextual keyword extraction with large language models. *arXiv preprint arXiv:2404.05406*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Minghui Qiu, Xinjing Huang, Cen Chen, Feng Ji, Chen Qu, Wei Wei, Jun Huang, and Yin Zhang. 2021. Reinforced history backtracking for conversational question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13718–13726.

Chen Qu, Liu Yang, Minghui Qiu, W Bruce Croft, Yongfeng Zhang, and Mohit Iyyer. 2019a. Bert with history answer embedding for conversational question answering. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, pages 1133–1136.

Chen Qu, Liu Yang, Minghui Qiu, Yongfeng Zhang, Cen Chen, W Bruce Croft, and Mohit Iyyer. 2019b. Attentive history selection for conversational question answering. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1391–1400.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training.

Muhammad Shihab Rashid, Jannat Ara Meem, and Vagelis Hristidis. 2024. Normy: Non-uniform history modeling for open retrieval conversational question answering. *arXiv preprint arXiv:2402.04548*.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Timo Schick and Hinrich Schütze. 2021. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352.

Yuan Sun, Chaofan Chen, Sisi Liu, and Xiaobing Zhao. 2021. Ti-reader: 基于注意力机制的藏文机器阅读理解端到端网络模型(ti-reader: An end-to-end network model based on attention mechanisms for tibetan machine reading comprehension). In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 219–228.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Svitlana Vakulenko, Shayne Longpre, Zhucheng Tu, and Raviteja Anantha. 2021. Question rewriting for conversational question answering. In *Proceedings of the 14th ACM international conference on web search and data mining*, pages 355–363.

Ellen M Voorhees, Dawn M Tice, et al. 1999. The trec-8 question answering track evaluation. In *TREC*, volume 1999, page 82.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Yi-Ting Yeh and Yun-Nung Chen. 2019. Flowdelta: Modeling flow information gain in reasoning for conversational machine comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 86–90.


(c) 2024 中国中文信息学会计算语言学专业委员会

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Wanjun Zhong, Yifan Gao, Ning Ding, Yujia Qin, Zhiyuan Liu, Ming Zhou, Jiahai Wang, Jian Yin, and Nan Duan. 2022. Proqa: Structural prompt-based pre-training for unified question answering. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4230–4243.

Chenguang Zhu, Michael Zeng, and Xuedong Huang. 2018. Sdnet: Contextualized attention-based deep network for conversational question answering. *arXiv preprint arXiv:1812.03593*.

Wenhao Zhuang, Yuan Sun, and Xiaobing Zhao. 2024. Tilamb：基于增量预训练的藏文大语言模型(tilamb: A tibetan large language model based on incremental pre-training). In *Proceedings of the 23th Chinese National Conference on Computational Linguistics*.

CCL 2024

# TiLamb：基于增量预训练的藏文大语言模型

庄文浩[1,2]　孙媛[1,2,3,*]　赵小兵[1,2,3]

[1]中央民族大学　信息工程学院，北京　100081

[2]国家语言资源监测与研究少数民族语言中心

[3]民族语言智能分析与安全治理教育部重点实验室

*通讯作者：孙媛

tracy.yuan.sun@gmail.com

## 摘要

基于“预训练+微调”范式的语言模型展现了卓越的性能，随着模型规模和训练数据量的扩增，其解决多种自然语言处理任务的能力得到了显著的提高。当前的大语言模型主要支持英汉等主流语言，这限制了藏语等低资源语言在该领域的研究。针对藏语数据稀缺、现有藏语预训练模型效果不够好、下游任务可扩展性差等问题，本文汇总清洗得到26.43GB藏文数据，以开源的LLaMA2-7B作为基座模型，扩充LLaMA2现有词表，增加了约30,000个藏文tokens，提高其藏文编码效率和对藏文的语义理解能力，通过增量预训练得到藏文大语言模型基座TiLamb。根据多种藏文下游任务分别制作数千到几万条不等的微调数据集，微调后的TiLamb在藏文新闻分类、藏文实体关系分类、藏文机器阅读理解、藏文分词、藏文摘要、藏文问题回答、藏文问题生成共七个下游任务中进行验证，多项指标结果相较传统方法和其他藏文预训练模型有大幅提升。本文将TiLamb和部分资源开放供研究使用，https://github.com/NLP-Learning/TiLamb。

**关键词：** 增量预训练；藏文大语言模型；指令微调；下游任务

# TiLamb: A Tibetan Large Language Model Based on Incremental Pre-training

**Wenhao Zhuang[1,2] Yuan Sun[1,2,3,*] Xiaobing Zhao[1,2,3]**

[1] School of Information Engineering, Minzu University of China, Beijing　100081

[2] National Language Resources Monitoring and Research Center for Minority Languages

[3] Key Laboratory of Ethnic Language Intelligent Analysis and Security Governance of MOE

*Corresponding author: Yuan Sun

tracy.yuan.sun@gmail.com

## Abstract

Based on the ”pre-training + fine-tuning” paradigm, language models have demonstrated exceptional performance. With the expansion of model size and training data volume, their ability to solve a variety of natural language processing tasks has seen significant improvements. Current large language models primarily support mainstream languages such as English and Chinese, limiting research in low-resource languages like Tibetan. Addressing issues such as the scarcity of Tibetan data, the inadequate performance of existing Tibetan pre-trained models, and poor expandability in downstream tasks, this paper consolidates and cleans 26.43GB of Tibetan data. Utilizing the open-source LLaMA2-7B as the foundational model, it expands the existing vocabulary of LLaMA2 by adding approximately 30,000 Tibetan tokens to enhance Tibetan text encoding efficiency and semantic understanding capabilities. Through incremental pre-training, the Tibetan large language model foundation, TiLamb, was developed.

第二十三届中国计算语言学大会论文集，第254页-第267页，太原，中国，2024年7月25日至28日。
卷1：主会论文
(c) 2024 中国中文信息学会计算语言学专业委员会

For various Tibetan downstream tasks, fine-tuning datasets ranging from thousands to tens of thousands of entries were prepared. The fine-tuned TiLamb was validated across seven Tibetan downstream tasks, including Tibetan news categorization, Tibetan entity relationship classification, Tibetan machine reading comprehension, Tibetan word segmentation, Tibetan summarization, Tibetan question answering, and Tibetan question generation, achieving significant improvements in multiple metrics over traditional methods and other Tibetan pre-trained models. This paper makes TiLamb and some resources available for research use at https://github.com/NLP-Learning/TiLamb.

**Keywords:** Incremental Pre-training , Tibetan Large Language Model , Prompt-based Fine-tuning , Downstream Tasks

## 1 引言

作为预训练语言模型（Pre-training Language Model, PLM）初期的创新性研究之一，ELMo(Sarzynska-Wawer et al., 2021)采用预训练的双向LSTM（BiLSTM）网络来实现对词汇在不同上下文中的动态表示，通过对特定的下游任务进行微调，进一步优化了BiLSTM网络的性能。近年来，通过在大规模语料库上对基于自注意力机制的高度并行化Transformer(Vaswani et al., 2017)架构进行预训练，人们发现BERT(Devlin et al., 2018)、GPT-2(Radford et al., 2019)、BART(Lewis et al., 2020)等基于"预训练+微调"范式的PLM在解决多种自然语言处理（Natural Language Processing, NLP）任务方面表现出强大的能力。进一步研究发现随着模型规模的扩大，这些语言模型的性能不仅得到了巨大提升，而且还表现出上述小规模语言模型所不具备的特殊能力（如上下文学习）。为区分不同参数规模下的语言模型，研究界创造了术语——大语言模型（Large Language Model, LLM）代指大型的PLM（包含数十亿乃至数千亿参数）。

有关藏语的预训练模型已经有了一些进展。Facebook AI的XLM-R模型(Conneau et al., 2020)使用了超过2TB的跨语言语料库，提出了一种高效、可扩展的跨语言表示学习方法，在近百种语言上进行了实验，取得了句子分类、命名实体识别等多个任务上的良好表现。哈工大讯飞联合实验室Yang等人(Yang et al., 2022)推出的少数民族语言的多语言预训练模型CINO是基于多语言预训练模型XLM-R开发的，涵盖了藏语等8种语言，在藏语新闻分类任务数据集TNCC(Qun et al., 2017)、朝鲜语新闻分类YNAT(Park et al., 2021)上获得了良好的效果。Liu等人(Liu et al., 2022)提出藏文预训练语言模型TiBERT，在问题生成等任务上取得突破。安波等人(安波and 龙从军, 2022)抓取了一个较大规模的藏文文本数据集，基于这些数据训练了BERT-base-Tibetan预训练模型，能够显著提升藏文文本分类的性能。Deng等人(Deng et al., 2023b)构建了包含蒙藏维哈朝五种少数民族语言的预训练模型MiLMo，为少数民族语言的各种下游任务提供支持。湖南大学Li等人(Li et al., 2022)使用超过10GB的汉英维藏蒙语料，在BART的基础上，加入DeepNorm训练得到少数民族预训练模型CMPT。Deng等人(Deng et al., 2023a)为了解决藏文预训练语言模型在知识记忆和推理能力上的缺陷，提出基于知识增强的藏文预训练语言模型TiKEM。

以上有关藏文的预训练模型虽然在多个下游任务上都取得了不错的表现，但均基于BERT设计，缺少解码器结构，并不属于真正的生成式藏文语言模型，且局限于分类等任务，模型参数量远低于LLM的参数量，对于藏文这种低资源语言，使用的训练数据还不够多。针对以上问题，本文提出基于LLaMA2-7B进行增量预训练的藏文大语言模型TiLamb（Tibetan Large Language Model Base），并在多个下游任务上验证了模型性能。本文的主要贡献如下：

（1）汇总清洗了26.43GB藏文预训练语料，使用SentencePiece(Kudo and Richardson, 2018)的BPE算法在近10GB藏文文本上训练得到词表大小为32,000的藏文分词模型。在原始LLaMA2的词表中增加额外约30,000个藏文tokens，增强了藏文的编码和解码效率，并提高了LLaMA2对藏文的理解能力。

（2）采用低秩适配（Low-Rank Adaptation, LoRA）方法，高效地对LLaMA2-7B进行增量预训练，并微调TiLamb适配藏文下游任务，在藏文新闻分类、藏文实体关系分类、藏文分

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

词、藏文摘要、藏文问题回答、藏文问题生成共六个下游任务中性能得到较大提升，在藏文机器阅读理解任务中取得较好结果。

（3）本文将藏文大语言模型TiLamb和部分资源公开，过程方法对藏语等低资源语言的LLM研究具有一定借鉴意义，促进低资源语言NLP的合作和发展。

## 2 相关工作

LLM的预训练对大规模文本数据的自动学习以捕获语言的复杂性和细粒度信息，能够在无需明确指示的情况下，捕获和理解文本中的隐含意义，从而增强其对新任务和未知数据的适应能力。LLM从零到一的预训练过程需要大规模的数据、计算资源以及时间，成本非个人和一般研究机构所能承受，因此，许多研究者利用现有的开源预训练模型作为基座，根据实际需求进行增量预训练和微调，往往能将大模型用于垂直领域，比如金融(Li et al., 2023)、法律(Blair-Stanek et al., 2023)等领域，并取得较好的效果。部分LLM训练统计数据如表1所示。

| 模型 | 参数大小 | 数据规模 | 计算资源 | 训练时间 |
|---|---|---|---|---|
| T5(Raffel et al., 2020) | 11B | 1万亿tokens | 1024 TPU v3 | - |
| LLaMA(Touvron et al., 2023a) | 65B | 1.4万亿tokens | 2048 80G A100 | 21天 |
| LLaMA2(Touvron et al., 2023b) | 70B | 2万亿tokens | 2048 80G A100 | 36天 |
| BLOOM(Le Scao et al., 2022) | 176B | 3600亿tokens | 384 80G A100 | 105天 |
| CodeGeeX(Zheng et al., 2023) | 13B | 8500亿tokens | 1536 Ascend 910 | 60天 |
| ERNIE3.0(Zhang et al., 2019) | 10B | 3750亿tokens | 384 V100 | - |
| FLAN(Wei et al., 2021) | 137B | - | 128 TPU v3 | 60小时 |
| Gopher(Rae et al., 2021) | 280B | 3000亿tokens | 4096 TPU v3 | 920小时 |

表 1. 部分LLM预训练统计数据

微调旨在通过在特定任务相关数据上进行有监督学习来调整预训练模型的参数(Wei et al., 2021)，通常涉及对模型的最后几层或所有层进行调整，以便模型能够更好地适应特定的数据集和任务。由于LLM包含大量的参数，参数高效微调是一个重要的课题，旨在减少可训练参数的数量，同时尽可能保持良好的性能。适配器微调在Transformer模型中引入了小型神经网络模块（称为适配器）(Houlsby et al., 2019)，适配器模块将被集成到每个Transformer层中，通常使用串行插入的方式，分别在Transformer层的两个核心部分（即注意力层和前馈层）之后。前缀微调(Li and Liang, 2021)在语言模型的每个Transformer层前添加了一系列前缀，这些前缀是一组可训练的连续向量。与前缀微调不同，提示微调(Lester et al., 2021)主要是在输入层中加入可训练的提示向量。低秩适配LoRA(Hu et al., 2021)通过添加低秩约束来近似每层的更新矩阵，以减少适配下游任务的可训练参数。以上多种微调方式可以在不重复昂贵预训练过程的同时，实现对模型的快速适应和优化。

预训练不同语种语料的占比影响着LLM对不同语言的学习能力。LLaMA预训练使用的语料中近90%是英文语料，中文语料占比不足1%，对中文理解和生成能力有所欠缺，为了提高其中文对话能力，需要加入中文数据进行增量预训练，之后再微调使其适应下游任务，许多开源的Github项目基于这种思想开发，如Llama-Chinese、Chinese-LLaMA-Alpaca1/2(Cui et al., 2023)、Chinese-Vicuna(Chenghao Fan and Tian, 2023)等。基于这种迁移学习的思想，使用藏文语料在开源模型上进行增量预训练和微调，有望让LLM理解和生成藏文。

在藏语方面，哈工大讯飞联合实验室(Yang et al., 2022)提出了少数民族多语言预训练模型CINO，其中包含0.13B tokens藏文，CINO-Large版本的参数量为0.585B，参与训练的多种少数民族语言的tokens总数为3.37B。Liu等人(Liu et al., 2022)构建了覆盖藏语语料库99.95%的词汇表，提出了藏文预训练语言模型TiBERT，参数量为0.11B，藏文语料约3.56GB。Deng等人(Deng et al., 2023a)针对藏文预训练语言模型缺少外部知识指导，知识记忆能力和知识推理能力受限的问题，使用含有50万个三元组知识的藏文知识增强预训练数据集，训练了基于知识增强的藏文预训练语言模型TiKEM，参数量为0.115B，参与训练tokens总数为0.245B。安波等


(c) 2024 中国中文信息学会计算语言学专业委员会

人(安波and 龙从军, 2022)抓取了较大规模的藏文文本数据集，基于这些数据训练了BERT-base-Tibetan预训练模型。现有的藏文预训练模型在藏文下游任务上虽然表现不错，但其可拓展性有限，结果仍有很大的提升空间。主要原因是模型规模小，训练数据不足。解决这些问题，提高藏文预训练模型的能力，是本文研究的重点和难点。

## 3 TiLamb藏文大语言模型

### 3.1 LLaMA2模型基础

LLaMA2系列模型(Touvron et al., 2023b)是基于transformer架构(Vaswani et al., 2017)的纯解码器网络，集成了预归一化(Zhang and Sennrich, 2019)、SwiGLU激活函数(Shazeer, 2020)和旋转位置编码(Su et al., 2024)，实现了模型性能与训练效率的优化。覆盖不同参数规模（7B至65B），LLaMA2通过混合数据源预训练，在语言建模任务上展现了卓越性能。该模型结构包括32层，每层采用自注意力机制，其中线性投影实现了查询、键、值和输出的转换，而旋转嵌入强化了序列位置的表征。模型的MLP部分采用门控单元，增加了处理能力的非线性，而归一化层保证了训练的稳定性。最终，语言模型头部（lm_head）将隐藏状态映射到与词汇表大小相当的输出空间。

### 3.2 构建藏文分词模型

LLaMA2原始的分词模型是基于SentencePiece库(Kudo and Richardson, 2018)训练的。SentencePiece是一个用于训练分词器的开源库，它支持多种训练算法，包括unigram、bpe、word和char。在训练过程中，LLaMA2使用了BPE（Byte Pair Encoding）算法，它通过合并常见的字节对来构建词汇表，从而能够有效地处理各种语言。

本文收集近10GB的原始藏文语料，使用SentencePiece训练得到词表大小为32,000、覆盖率为99.95%的藏文分词模型，重要参数选择如表2所示，其余参数均为默认。

为了优化和适配LLaMA2的词表设计，本文在模型训练时进行了多项策略优化。首先，将所有数字字符拆分为单独的单元，减少了词表的数字量。其次，开启字节回退，在遇到未知或少见字符时将其分解为UTF-8字节来表示，避免OOV问题。此外，除了标准的藏文语料外，还加入了藏文的方言和变体（如古藏语），以确保模型能够处理各种地域性语言特征，提高分词模型的鲁棒性。

| 参数 | 值 | 参数 | 值 |
|---|---|---|---|
| 词表大小(`vocab_size`) | 32000 | 分词算法(`model_type`) | BPE |
| 数字拆分(`split_digits`) | True | 字节回退(`byte_fallback`) | True |
| 最大句子长度(`max_sentence_length`) | 5000 | 字符覆盖率(`character_coverage`) | 0.9995 |

表 2. 藏文分词模型训练参数

### 3.3 LLaMA2词表藏文扩充

LLaMA2-7B的预训练语料包含约2万亿个tokens，其中将近90%是英语，少部分是使用拉丁或西里尔字母的其他欧洲语言，因此LLaMA2具备多语言和跨语言理解能力。

本文在对LLaMA2进行初步研究时发现两个问题：（1）其对藏文的理解能力很差，无法理解藏文指令，这个问题的原因是LLaMA2的预训练语料缺少藏文，或者藏文语料占比太小导致预训练过程无法有效关注到藏文的向量表示；（2）LLaMA2在对藏文指令回复时，大部分是用英文回复无法理解指令，偶有的藏文回复也是对问题的重复或者错乱的藏文字符组合，且推理速度相比英文问答极其缓慢，导致速度缓慢的原因是LLaMA2词表中几乎没有藏文字符，尽管LLaMA2的分词器通过将未在词表中的内容分成多个Unicode字符来规避问题，但这样显著增加了序列的长度，而且降低了藏文文本的编码和解码效率。此外，每个藏文字符被分割为若干个Unicode字符，这些Unicode字符也可以表示其他语言，这会破坏藏文语义和语法结构的完整性，从而使得LLaMA2很难有效地学习捕捉藏文字符的语义表示。


(c) 2024 中国中文信息学会计算语言学专业委员会

本文将**3.2**节训练得到的藏文词表合并到LLaMA2原始词表中，合并后的词表大小为61,221，称为TiLamb分词器。对于同样一句藏文，使用LLaMA2原始分词器和TiLamb分词器的对比如表3所示。

| | 长度 | 内容 |
|---|---|---|
| 藏文语句 | 30 | ཅ་དངོས་བླུགས་སྒམ་ཁ་ཕྱེ་ནས་ཞིབ་བཤེར་བྱས་རྗེས། འགག་སྒོ་ལས་ཁུངས་ཀྱི་འབྲེལ་ཡོད་མི་སྣས་ཐུམ་སྒམ་ནང་ཇའི་སྒམ་ཆུང་ཞིག་རྙེད་ཅིང་། （打开行李箱检查后，海关人员在包装箱内发现了一个茶盒。） |
| LLaMA2 分词器 | 174 | '▁', '<0xE0>', '<0xBD>', '<0x85>', '·', 'ད', 'ང', 'ོ', 'ས', '·', 'བ', '<0xE0>', '<0xBE>', '<0xB3>', 'ུ', 'ག', 'ས', '·', 'ས', '<0xE0>', '<0xBE>', '<0x92>', 'མ', '·', ••• 省略剩余的 150 个 tokens |
| TiLamb 分词器 | 19 | '▁ཅ་', 'དངོས་', 'བླུགས་', 'སྒམ་', 'ཁ་ཕྱེ་ནས་', 'ཞིབ་བཤེར་', 'བྱས་རྗེས།', '▁', 'འགག་སྒོ་', 'ལས་ཁུངས་ཀྱི་', 'འབྲེལ་ཡོད་', 'མི་སྣས་', 'ཐུམ་', 'སྒམ་', 'ནང་', 'ཇའི་', 'སྒམ་ཆུང་', 'ཞིག་རྙེད་', 'ཅིང་།' |

表 3. LLaMA2原始分词器和TiLamb分词器的分词结果对比

实验对比表明，对于相同一段藏文文本，换用TiLamb分词器后token数量从174缩减至19，减少约8倍，大幅提高了模型对藏文的编码效率和推理速度。在固定的上下文长度下，模型的最大输入/输出文本长度、模型可以容纳的信息量均提升至原来的8倍以上。

### 3.4 扩展词嵌入和调整lm_head维度

为了适配TiLamb分词器，本文将词嵌入和语言模型头部从原始尺寸$V \times H$调整为新尺寸$V' \times H$，其中$V = 32,000$代表原始词汇表大小，$V' = 61,221$是TiLamb分词器的词表大小。

LLaMA2-7B原始的Embedding层的大小为$(32000, 4096)$，即词表里的每一个token对应一个$1 \times 4096$的Embedding向量。将新增的token在模型的Embedding层和lm_head层进行初始化，常用的有均值扩充和随机扩充两种方法，本文选用业界常用的均值扩充方法，即新增token的Embedding用原来token的Embedding的均值来表示。新的藏文词表被添加到原始嵌入矩阵的末尾，确保原始词汇表中的词嵌入不受影响。初始化后可得新增训练参数量为$(61221 - 32000) \times 4096 \times 2 = 239,378,432$。

### 3.5 预训练数据收集与处理

本文预训练数据主要包括：藏文新闻分类数据集TNCC(Qun et al., 2017)、UTibetNLP分类的藏文新闻数据(Zhang et al., 2022)、早期积累的藏文语料约15.5GB。爬取人民网藏文版、云藏网、西藏新闻网等二十余个藏文网站共计约7.5GB网页数据，包括党政文件、通知公告等多个领域。爬取云藏百科、维基百科等藏文百科数据3.1GB。爬取小红书和微信公众号文章约0.3GB。

以上数据均经过去重、隐私去除、质量过滤三种方式处理：

（1）去重：语料中的重复数据会降低语言模型的多样性，可能导致训练过程不稳定，从而影响模型性能(Hernandez et al., 2022),因此本文直接对重复句子进行过滤，同时删除包含重复单词或短语的低质量语句。

（2）隐私去除：本文的预训练文本数据大都来源于网络，包含涉及敏感或个人信息的内容，如联系方式、地址、邮箱等，这会增加隐私泄露的风险(Carlini et al., 2021)。本文使用直接有效的基于规则的方法，包括关键字识别、正则匹配等方法来对隐私内容进行去除。

（3）质量过滤：本文结合特定关键词的集合以及正则匹配，识别并删除文本中的噪声和无用元素，包括一些HTML标签、超链接、模板和攻击性词语。

处理后的藏文无标签预训练数据大小为26.43GB，藏文tokens总数约3B。

### 3.6 使用LoRA进行参数高效微调

传统的全参数更新训练范式对于大型语言模型来说计算时间和成本过高，低秩适配LoRA(Hu et al., 2021)是一种参数高效的训练方法，它保持预训练模型权重不变，同时引入可训练的秩分解矩阵。LoRA冻结了预训练模型的权重，并在每一层注入可训练的低秩矩


(c) 2024 中国中文信息学会计算语言学专业委员会

阵。具体来说，对于一个线性层，其权重矩阵为$W_0 \in \mathbb{R}^{d\times k}$，其中$k$是输入维度，$d$是输出维度，LoRA添加了两个低秩分解的可训练矩阵$B \in \mathbb{R}^{d\times r}$和$A \in \mathbb{R}^{r\times k}$，其中$r$是预定的秩。带有输入$x$的前向传播由公式(1)给出。

$$h = W_0x + \Delta Wx = W_0x + BAx, \quad B \in \mathbb{R}^{d\times r}, A \in \mathbb{R}^{r\times d} \tag{1}$$

在训练过程中，$W_0$ 是固定的，不接收梯度更新，而$B$ 和$A$ 是可更新的。通过选择$r \ll \min(d, k)$，能够减少显存消耗。

为在成本和性能之间选择平衡点，实现参数的高效训练，本文将LoRA训练方法应用于增量预训练和微调两个阶段，主要将LoRA适配器集成到注意力模块和MLP层的权重中。

### 3.7 预训练目标

本文使用26.43GB藏文预训练数据在标准的因果语言模型（Causal Language Modeling, CLM）任务上对LLaMA2模型进行增量预训练。给定一个输入token序列$x = (x_0, x_1, x_2, \ldots)$，模型被训练以自回归的方式预测下一个token $x_i$，目标是最小化负对数似然，如公式(2)所示。

$$\mathcal{L}_{\text{CLM}}(\Theta) = \mathbb{E}_{x\sim D_{PT}}\left[-\sum_i \log p(x_i|x_0, x_1, \ldots, x_{i-1}; \Theta)\right] \tag{2}$$

其中，$\Theta$ 表示模型参数，$D_{PT}$ 是预训练数据集，$x_i$ 是待预测的token，$x_0, x_1, \ldots, x_{i-1}$ 构成上下文。

本文使用LLaMA-Factory框架(Zheng et al., 2024)进行LLaMA2-7B的增量预训练，部分参数的选择如表4所示。增量预训练之后得到藏文大语言模型TiLamb，过程中损失与学习率的记录如图1所示。

| 参数 | 值 | 参数 | 值 |
|---|---|---|---|
| cutoff_len | 1024 | learning_rate | $2 \times 10^{-4}$ |
| finetuning_type | lora | num_train_epochs | 1.0 |
| per_device_train_batch_size | 4 | gradient_accumulation_steps | 2 |
| lr_scheduler_type | cosine | max_grad_norm | 1.0 |
| lora_rank | 8 | lora_dropout | 0.1 |
| lora_target | q_proj, v_proj | warmup_steps | 0 |
| additional_target | embed_tokens, lm_head, norm | fp16 | True |

表 4. 增量预训练过程参数选择

### 3.8 TiLamb监督微调

预训练语言模型难以遵循用户指令，经常会生成非预期内容。这是因为公式(2)中的建模目标是预测下一个token，而不是根据指令回答问题(Ouyang et al., 2022)。为了使语言模型的行为与用户的意图对齐，可以通过微调模型明确训练它遵循指令。本文采用LLaMA-Factory框架中的微调模板，其中instruction为用于描述任务的指令，input可选，为任务指令的补充输入，output为期望模型产生的回复，示例如图2所示。

在监督微调过程中，损失仅在输入序列的“output”部分计算，如公式(3)。

$$\mathcal{L}_{\text{SFT}}(\Theta) = \mathbb{E}_{x\sim D_{\text{SFT}}}\left[-\sum_{i\in output} \log p(x_i|x_0, x_1, \ldots, x_{i-1}; \Theta)\right] \tag{3}$$

其中，$\Theta$ 表示模型参数，$D_{\text{SFT}}$ 是微调数据集，$x = (x_0, x_1, \ldots)$ 代表分词后的输入序列。


(c) 2024 中国中文信息学会计算语言学专业委员会

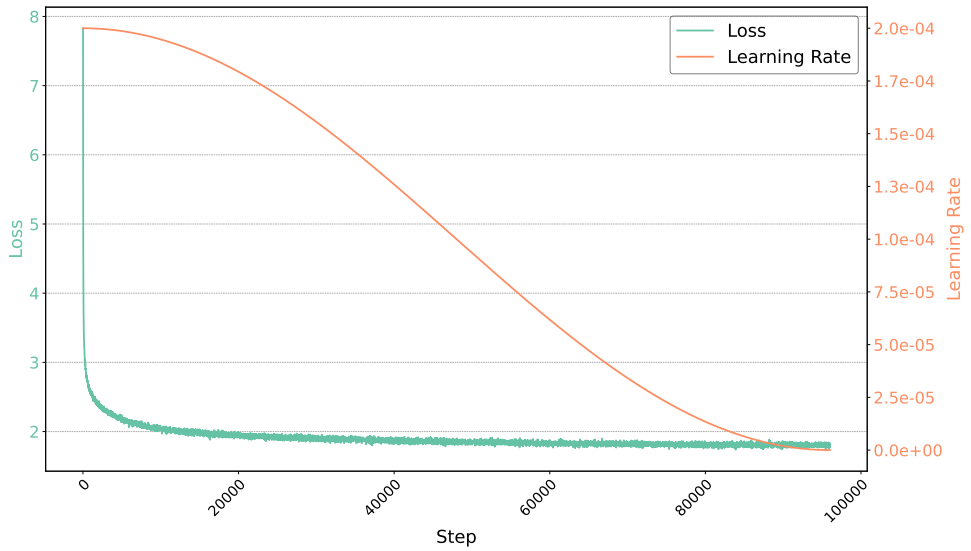



图 1. 增量预训练过程Loss与Learning Rate的变化

"instruction": "AIལས་རོགས་པ་ལ་མི་ཞེག་གིས་འཚམས་འདྲི་ཕྱིད་དུས་ལན་འདེབས་ཚོག་པའི་འཚམས་འདྲི་ཞུ་ཚོག་ཅིག་གསར་འཕྲུག་ཐུབ།"（创建一个 AI 助手在被人问候时可以说的问候语。）",

"input": "",

"output": "སྐུ་འཚས་བཟང་། དེ་རིང་ངས་ཁྱོད་ལ་ཅི་ཞེག་བསྒྲུབ་ཐུབ།"（“你好！今天我能帮你做些什么？”）"

图 2. 微调模板示例

## 4 TiLamb下游任务实验评估

在藏文大语言模型基座TiLamb上针对各项下游任务进行LoRA微调，合并适配器后调用API进行自动测试，收集实验结果并进行科学分析。TiLamb在藏文下游任务的监督微调中部分参数的选择如表5所示。

本文采用“一种下游任务对应一个微调模型”的策略，主要考虑到各类下游任务的数据规模和难度差异较大，使大模型难以同时掌握多种任务。此外，初步实验表明，设置不同提示词（prompt）虽能帮助模型区分任务，但会降低单个任务的性能。该策略无需在微调时设置提示词，提升了训练效率，并简化了API调用和用户使用过程。

| 参数 | 值 | 参数 | 值 |
|---|---|---|---|
| cutoff_len | 2048 | learning_rate | $2 \times 10^{-4}$ |
| finetuning_type | lora | num_train_epochs | 3.0 |
| per_device_train_batch_size | 4 | gradient_accumulation_steps | 2 |
| lr_scheduler_type | cosine | max_grad_norm | 1.0 |
| lora_rank | 8 | lora_dropout | 0.05 |
| lora_target | q_proj, v_proj | fp16 | True |

表 5. 监督微调过程参数选择

### 4.1 藏文新闻分类

该任务选用由复旦大学自然语言处理实验室发布的藏语新闻数据集Tibetan News Classifi-


(c) 2024 中国中文信息学会计算语言学专业委员会

cation Corpus (TNCC)(Qun et al., 2017)。数据集包含9,204条样本，涉及政治、经济、教育、旅游、环境、艺术、文学、宗教等12个类别。本文按9:1的比例将其划分为训练集和测试集，训练集的数据用于制作微调数据集。评价指标为Accuracy(%)和Macro-F1(%)，实验结果如表6所示。

| 模型 | Accuracy(%) | Macro-F1(%) |
|---|---|---|
| Transformer(Vaswani et al., 2017) | 28.63 | 28.79 |
| CNN(syllable) | 61.51 | 57.34 |
| TextCNN(Guo et al., 2019) | 61.71 | 61.53 |
| DPCNN(Johnson and Zhang, 2017) | 62.91 | 61.17 |
| TextRCNN(Lai et al., 2015) | 63.67 | 62.81 |
| Bert-base-Tibetan(安波and 龙从军, 2022) | - | 51.00 |
| TiBERT(Liu et al., 2022) | 71.04 | 70.94 |
| CINO-base(Yang et al., 2022) | 73.10 | 70.00 |
| TiKEM(Deng et al., 2023a) | 74.46 | 72.61 |
| **TiLamb+LoRA** | **78.85** | **77.45** |

表 6. 藏文新闻文本分类结果

TiLamb使用训练集经LoRA微调后，在藏语新闻文本分类上取得了最好的分类效果，准确率比CINO-base和TiKEM分别提高了5.75%和4.39%，F1值比CINO-base和TiKEM分别提高了7.45%和4.84%。

### 4.2 藏文实体关系分类

为了验证TiLamb模型对知识的记忆及融合运用能力，本文使用6,433条三元组-文本对齐数据集，三元组中共有11种关系。该任务要求在给定两个实体和包含该实体的对应文本后，给出两个实体之间的关系类别。本文按9:1的比例将其划分为训练集和测试集，训练集的数据用于制作该任务的微调数据集。评价指标为Accuracy(%)、Macro-P(%)、Macro-R(%)和Macro-F1(%)，实验结果如表7所示。

| 模型 | Accuracy(%) | Macro-P(%) | Macro-R(%) | Macro-F1(%) |
|---|---|---|---|---|
| FastText(Joulin et al., 2016) | 55.80 | 34.05 | 32.98 | 31.61 |
| DPCNN(Johnson and Zhang, 2017) | 70.94 | 54.21 | 49.23 | 48.65 |
| TextCNN(Guo et al., 2019) | 72.38 | 71.03 | 59.11 | 56.76 |
| TiBERT(Liu et al., 2022) | 84.70 | 76.66 | 68.82 | 67.94 |
| CINO-base(Yang et al., 2022) | 85.31 | 75.48 | 69.12 | 66.73 |
| MiLMO(Deng et al., 2023b) | 85.76 | 77.13 | 68.97 | 68.57 |
| TiKEM(Deng et al., 2023a) | 90.12 | 91.73 | 75.61 | 76.34 |
| **TiLamb+LoRA** | **95.98** | **97.14** | **88.98** | **91.60** |

表 7. 藏文实体关系分类结果

TiLamb微调后在该任务上的分类准确率（Accuracy）上达到了95.98%，相比于其他预训练模型，TiLamb在所有评价指标上都有明显提升。


(c) 2024 中国中文信息学会计算语言学专业委员会

### 4.3 藏文机器阅读理解

机器阅读理解任务是给定一段文本和一个问题，让模型回答对应问题。这需要模型理解问题和上下文语义，然后进行推理、判断等，给出具体答案。本文使用藏文机器阅读理解数据集TibetanQA(Sun et al., 2021a)对模型的阅读理解能力进行评估，该数据集包含了1,513篇文章和20,000个问答对。为了评估模型性能，本文使用EM值（精确匹配）和F1值作为评价指标。

本文以8:2的比例将数据划分为训练集和测试集，训练集用于制作该任务的微调数据集，实验结果如表8所示。

| 模型 | EM(%) | F1(%) |
|---|---|---|
| R-Net(Wang et al., 2017) | 55.8 | 63.4 |
| BiDAF(Seo et al., 2016) | 58.6 | 67.8 |
| QANet(Yu et al., 2018) | 57.1 | 66.9 |
| TiBERT(Liu et al., 2022) | 53.2 | 73.4 |
| **TiLamb+LoRA** | 46.6 | 77.4 |
| Ti-Reader(Sun et al., 2021c) | 67.9 | 77.4 |
| TiKEM(Deng et al., 2023a) | **69.4** | **80.1** |

表 8. 在TibetanQA上的藏文阅读理解评估结果

TiLamb作为生成式模型，在做抽取式阅读理解任务时相比其他模型难度大，因为其他模型只需根据文本和问题，在文本中找出答案的起止位置即可，但生成式模型本就基于概率生成。即便如此，微调后的TiLamb的F1值依然达到了77.4%，与藏文抽取式机器阅读理解模型Ti-Reader的指标相同，只比TiKEM的F1值低2.7%。

### 4.4 藏文分词

对于藏文这种结构复杂、资源相对较少的语言而言，正确的分词对于进一步的语言处理任务，如语义分析、机器翻译和信息检索等，都具有至关重要的作用。该任务使用中文信息学会举办的第一届藏文分词评测所用的数据集，数据集中藏语短句去重后有21,427条，本文保留了1,000条作为测试集，剩余的20,427条用于制作微调数据集，测试结果与第一届藏文分词评测第一名TIP-LAS(李亚超et al., 2015)对比如表9所示。

| 模型 | Precision(%) | Recall(%) | F1(%) |
|---|---|---|---|
| TIP-LAS(李亚超et al., 2015) | 93.14 | 92.17 | 92.66 |
| **TiLamb+LoRA** | **93.58** | **93.71** | **93.64** |

表 9. 藏文短句分词评估结果

实验结果表明，TiLamb微调后在藏文短句分词任务中表现出色，其Precision、Recall 和F1值均超过了TIP-LAS 模型。具体而言，微调后的TiLamb的F1 值达到93.64%，比TIP-LAS提高了0.98个百分点。

### 4.5 藏文摘要

文本摘要是自然语言处理领域中的一项关键技术，它通过对大量文本进行提炼和压缩，去除冗余信息，生成简洁而全面的概要，从而帮助用户迅速把握文本的核心内容，大幅提升阅读效率。本文使用47,088条新闻与对应摘要做微调，测试集使用5,232条新闻与对应摘要。用于本文实验的数据均以新闻标题作为摘要，这种方法能够很好地捕捉新闻的主要内容，确保生成的摘要简明扼要且信息丰富。实验结果如表10所示。


(c) 2024 中国中文信息学会计算语言学专业委员会

| 模型 | ROUGE-1(%) | ROUGE-2(%) | ROUGE-L(%) |
|---|---|---|---|
| 统一模型(Huang et al., 2023) | 19.81 | 13.27 | 16.90 |
| CMPT模型（Ti-SUM）(Huang et al., 2023) | 39.53 | 26.42 | 38.02 |
| CMPT（50000条）(Huang et al., 2023) | 49.16 | 33.43 | 48.66 |
| **TiLamb+LoRA** | **53.99** | **37.22** | **52.89** |

表 10. 藏文摘要生成评估结果

### 4.6 藏文问题回答

该任务使用本组人工制作的TiconvQA藏文多轮对话数据集，其中包含2,120条藏文文章段落及20,420对多轮问答对。为了进行实验评估，本文按照8:2的比例将数据集划分为训练集和测试集。在评估实验中，使用EM值和F1值作为评估指标，实验结果如表11所示。

| 模型 | EM(%) | F1(%) |
|---|---|---|
| DrQA(Chen et al., 2017) | 41.49 | 61.51 |
| TiBERT(Liu et al., 2022) | 45.12 | 65.71 |
| **TiLamb+LoRA** | **45.28** | **72.84** |

表 11. 单轮藏文问题问答评估结果

### 4.7 藏文问题生成

问题生成是自然语言生成的一项任务，它以文本和目标答案为输入，自动从答案中生成问题。本文使用用于机器阅读理解的藏文问答数据集TibetanQA(Sun et al., 2021a)，按照约9:1的比例划分训练集和测试集，其中用于微调的数据共17,762条，用于测试的数据为1,976条，实验指标数值均为百分比，实验结果如表12所示。

| 模型 | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | ROUGE-L |
|---|---|---|---|---|---|
| S2S+ATT+CP(Sun et al., 2021b) | 29.99 | 20.14 | 13.90 | 9.59 | 31.45 |
| TiBERT(Liu et al., 2022) | 35.48 | 28.60 | 24.51 | 21.30 | 40.04 |
| TiBERT+wh(Sun et al., 2022) | 42.45 | 35.07 | **29.64** | **25.58** | 43.28 |
| **TiLamb+LoRA** | **44.60** | **35.24** | 28.88 | 24.47 | **50.42** |

表 12. 藏文问题生成评估结果

微调后的TiLamb在问题生成任务中有不错的表现，尤其在ROUGE-L上达到了50.42%，相较其他模型有较大提升。

## 5 总结与展望

本文使用近10GB藏文数据训练得到词表大小为32,000的藏文分词模型，并用约30,000个藏文tokens扩充了LLaMA2现有词表，显著提高了其对藏文的编码效率，收集整理了26.43GB藏文预训练语料，并对LLaMA2-7B进行增量预训练得到藏文大语言模型TiLamb。针对七个藏文NLP下游任务分别微调后的TiLamb，经实验验证相较其他模型均有不同程度的提升。本文的过程和方法对其他低资源语言在大型语言模型方向的研究提供了一定的参考价值。当前的TiLamb仍存在一定局限性，由于预训练语料大部分为藏文新闻，对TiLamb进行百万数量级的指令微调后，虽能显著提升LLaMA2的藏文对话能力和指令遵循能力，但TiLamb的藏文理解


(c) 2024 中国中文信息学会计算语言学专业委员会

能力还有一定进步空间。在未来将使用更多有关藏族文化、历史和宗教的相关数据在参数量更大的LLM上进行预训练和通用对话微调，并加入人类反馈的强化学习（RLHF）过程，使模型的输出与人类偏好对齐，同时探索DoRA(Liu et al., 2024)、GaLore(Zhao et al., 2024)等更先进的微调算法对藏文大语言模型的适用性。

## 致谢

本论文得到了国家社科基金(22&ZD035), 国家自然科学基金(61972436), 中央民族大学项目(GRSCP202316, 2023QNYL22, 2024GJYY43)的资助。

## 参考文献

Andrew Blair-Stanek, Nils Holzenberger, and Benjamin Van Durme. 2023. Can gpt-3 perform statutory reasoning? In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, pages 22–31.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879.

Zhenyi Lu Chenghao Fan and Jie Tian. 2023. Chinese-vicuna: A chinese instruction-following llama-based model.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

Junjie Deng, Long Chen, Yan Zhang, Yuan Sun, and Xiaobin Zhao. 2023a. Tikem: 基于知识增强的藏文预训练语言模型(tikem: Knowledge enhanced tibetan pre-trained language model). In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics*, pages 135–144.

Junjie Deng, Hanru Shi, Xinhe Yu, Wugedele Bao, Yuan Sun, and Xiaobing Zhao. 2023b. Milmo: minority multilingual pre-trained language model. In *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 329–334. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Bao Guo, Chunxia Zhang, Junmin Liu, and Xiaoyi Ma. 2019. Improving text classification with weighted word embeddings via a multi-channel textcnn model. *Neurocomputing*, 363:366–374.

Danny Hernandez, Tom Brown, Tom Conerly, Nova DasSarma, Dawn Drain, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Tom Henighan, Tristan Hume, et al. 2022. Scaling laws and interpretability of learning from repeated data. *arXiv preprint arXiv:2205.10487*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Shuo Huang, Xiaodong Yan, Xinpeng OuYang, and Jinpeng Yang. 2023. 基于端到端预训练模型的藏文生成式文本摘要(abstractive summarization of tibetan based on end-to-end pre-trained model). In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics*, pages 113–123.


(c) 2024 中国中文信息学会计算语言学专业委员会

Rie Johnson and Tong Zhang. 2017. Deep pyramid convolutional neural networks for text categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 562–570.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71.

Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597.

Bin Li, Yixuan Weng, Bin Sun, and Shutao Li. 2022. A multi-tasking and multi-stage chinese minority pre-trained language model. In *China Conference on Machine Translation*, pages 93–105. Springer.

Jiangtong Li, Yuxuan Bian, Guoxuan Wang, Yang Lei, Dawei Cheng, Zhijun Ding, and Changjun Jiang. 2023. Cfgpt: Chinese financial assistant with large language model. *arXiv preprint arXiv:2309.10654*.

Sisi Liu, Junjie Deng, Yuan Sun, and Xiaobing Zhao. 2022. Tibert: Tibetan pre-trained language model. In *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2956–2961. IEEE.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024. Dora: Weight-decomposed low-rank adaptation. *arXiv preprint arXiv:2402.09353*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Sungjoon Park, Sungdong Kim, Jihyung Moon, Won Ik Cho, Kyunghyun Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Yongsook Song, et al. 2021. Klue: Korean language understanding evaluation. In *Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS 2021)*. Advances in Neural Information Processing Systems.

Nuo Qun, Xing Li, Xipeng Qiu, and Xuanjing Huang. 2017. End-to-end neural text classification for tibetan. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data: 16th China National Conference, CCL 2017, and 5th International Symposium, NLP-NABD 2017, Nanjing, China, October 13-15, 2017, Proceedings 5*, pages 472–480. Springer.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.


(c) 2024 中国中文信息学会计算语言学专业委员会

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Justyna Sarzynska-Wawer, Aleksander Wawer, Aleksandra Pawlak, Julia Szymanowska, Izabela Stefaniak, Michal Jarkiewicz, and Lukasz Okruszek. 2021. Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Research*, 304:114135.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. In *International Conference on Learning Representations*.

Noam Shazeer. 2020. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Y Sun, S Liu, C Chen, Z Dan, and X Zhao. 2021a. Construction of high-quality tibetan dataset for machine reading comprehension. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 208–218.

Yuan Sun, Chaofan Chen, Andong Chen, and Xiaobing Zhao. 2021b. Tibetan question generation based on sequence to sequence model. *Computers, Materials & Continua*, 68(3).

Yuan Sun, Chaofan Chen, Sisi Liu, and Xiaobing Zhao. 2021c. Ti-reader: 基于注意力机制的藏文机器阅读理解端到端网络模型(ti-reader: An end-to-end network model based on attention mechanisms for tibetan machine reading comprehension). In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 219–228.

Yuan Sun, Sisi Liu, Zhengcuo Dan, and Xiaobing Zhao. 2022. Question generation based on grammar knowledge and fine-grained classification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6457–6467.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. 2017. Gated self-matching networks for reading comprehension and question answering. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 189–198, Vancouver, Canada, July. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Ziqing Yang, Zihang Xu, Yiming Cui, Baoxin Wang, Min Lin, Dayong Wu, and Zhigang Chen. 2022. Cino: A chinese minority pre-trained language model. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3937–3949.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. In *International Conference on Learning Representations*.

Biao Zhang and Rico Sennrich. 2019. Root mean square layer normalization. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 12381–12392.

CCL 2024

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. Ernie: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451.

Jiangyan Zhang, Deji Kazhuo, Luosang Gadeng, Nyima Trashi, and Nuo Qun. 2022. Research and application of tibetan pre-training language model based on bert. In *Proceedings of the 2022 2nd International Conference on Control and Intelligent Robotics*, ICCIR '22, page 519–524, New York, NY, USA. Association for Computing Machinery.

Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. 2024. Galore: Memory-efficient llm training by gradient low-rank projection. *arXiv preprint arXiv:2403.03507*.

Qinkai Zheng, Xiao Xia, Xu Zou, Yuxiao Dong, Shan Wang, Yufei Xue, Zihan Wang, Lei Shen, Andi Wang, Yang Li, et al. 2023. Codegeex: A pre-trained model for code generation with multilingual evaluations on humaneval-x. *arXiv preprint arXiv:2303.17568*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*.

安波and 龙从军. 2022. 基于预训练语言模型的藏文文本分类. 中文信息学报, 36(12):85–93.

李亚超, 江静, 加羊吉, and 于洪志. 2015. Tip-las: 一个开源的藏文分词词性标注系统. 中文信息学报, 29(6):203–207.

CCL 2024

# 基于蒙古文文本语义辅助的噪声鲁棒蒙古语语音情感识别方法研究

**刘欢，梁凯麟，左昊麟，刘瑞 ***
内蒙古大学计算机学院，内蒙古，中国
happinessmonster@163.com, liangkailin98@foxmail.com
zuohaolin_0613@163.com, liurui_imu@163.com

## 摘要

噪声环境下语音情感识别（Speech Emotion Recognition，SER）旨在从带有背景噪声的语音信号中挖掘情感特征并自动预测说话人的情感状态。尽管这项技术在英语、汉语等语言方面取得了迅速的进展，但对于像蒙古语这样的小语种，在噪声环境下的语音情感识别研究仍处于起步阶段，缺乏相关数据集和方法的研究。为了推动蒙古语语音情感识别的发展，本研究首先构建了一个单说话人语音情感识别数据集。之后为了实现噪声环境下准确的蒙古语语音情感识别，我们提出了一种基于文本-语音双模态的带噪蒙古语语音情感识别基线模型 MonSER。文本信息为噪声语音信号提供额外的语义信息。具体来说，我们的模型首先对带噪语音信号进行频谱特征提取，之后使用多语种预训练模型 XLMBert 对语音信号对应的蒙古文文本信息进行编码。随后将上述提取的双模态信息进行融合，并输入分类器进行情感类别的预测。我们利用该数据集进行模型训练并测试模型的有效性。实验结果表明，我们的双模态模型在多种噪声环境下的蒙古语语音情感识别准确率明显优于只以语音为输入的单模态语音情感识别系统。同时，为了模拟实际场景中文本可能缺失的情况，我们提出了两种文本 mask 策略，该文本实验也进一步验证了文本语音双模态的有效性。

**关键词：** 蒙古语语音情感识别；噪声；文本

# Research on Noise-Robust Mongolian Speech Emotion Recognition Methods Based on Mongolian Text Semantics

**Huan Liu, Kailin Liang, Haolin Zuo, Rui Liu ***
Inner Mongolia University, School of Computer Science, Inner Mongolia, China
happinessmonster@163.com, liangkailin98@foxmail.com
zuohaolin_0613@163.com, liurui_imu@163.com

## Abstract

Speech Emotion Recognition (SER) in noisy environments aims to extract emotional features from speech signals with background noise and automatically predict the

* 通信作者：刘瑞（liurui_imu@163.com） ©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

speaker's emotional state. While this technology has rapidly advanced in major languages like English and Chinese, research on SER in minor languages such as Mongolian is still in its early stages, lacking relevant datasets and methodologies.To advance Mongolian Speech Emotion Recognition (SER), this study initially constructed a single-speaker SER dataset. Subsequently, to achieve accurate Mongolian SER in noisy environments, we proposed a dual-modality Text-Audio-based Noisy Mongolian Speech Emotion Recognition baseline model (MonSER).Textual information provides additional semantic information for noisy speech signals. Specifically, our model first extracts spectrogram features from noisy speech signals, then encodes the corresponding Mongolian text information using the multilingual pre-trained model XLMBert. The extracted dual-modal information is then fused and inputted into a classifier for emotional category prediction. We trained the model using this dataset and tested its effectiveness.Experimental results demonstrate that our dual-modality model achieves significantly higher accuracy in Mongolian Speech Emotion Recognition across various noise environments compared to single-modal speech emotion recognition systems that rely solely on speech inputs. Furthermore, to simulate the potential absence of text in real-world scenarios, we proposed two text masking strategies. This textual experiment also further validates the effectiveness of the text-speech bimodal approach.

**Keywords:** Mongolian Speech Emotion Recognition , Noise , Text

## 1 简介

带噪语音情感识别 (Speech Emotion Recognition, SER) 是一项旨在从背景噪声中的语音信号提取情感特征并预测说话人情感状态的技术。不同于传统的语音情感识别任务，带噪 SER 从现实应用的角度出发 (Win and Khine, 2020)，模拟了真实世界中的嘈杂环境。其目标在于在复杂的噪声环境中准确识别用户情感的状态 (Nam and Lee, 2021)(Tiwari et al., 2020)，从而帮助机器更好地理解用户的情感需求，为用户提供更加智能和贴心的服务 (Chatterjee et al., 2021)。它在各种人机交互场景 (Chenchah, Lachiri, 2016)(Alnuaim et al., 2022) 中具有巨大的应用潜力。例如，智能助手 (Xu, 2019)、客服机器人 (Bi et al., 2022)(Lee et al., 2020)、情感驾驶辅助系统 (Haghighat et al., 2023)、远程教育与会议 (Barron-Estrada et al., 2019)(Wang et al., 2020)(El Hammoumi et al., 2018)、健康监测和辅助诊断 (Singh and Srivastava, 2023)(Zisad et al., 2020) 等领域中。因此，这一领域的研究具有重要的意义和价值。

在现有的带噪语音情感识别研究中，面临着两个关键的挑战：1）任务相关的数据集支持；2）噪声环境下鲁棒的情感识别方法。为了更好地支持带噪 SER 的研究，先前的研究人员在数据集和方法论方面进行了大量的工作。在数据集方面，主流语言的情感语音库较为丰富，如 Maribor(Ambrus, 2000)、Belfast(McGilloway, 2000)(Douglas-Cowie, 2000)、EMO-DB(Burkhardt et al., 2005)、CASIA(Liu et al., 2011)、IEMOCAP(Busso et al., 2008) 等数据集，涵盖了英语、斯洛文尼亚语、法语和西班牙语、德语、中文等多种语言，情感种类也从 5 类情感到 8 类情感不等；然而，这些数据集通常只提供情感标签，没有对音频中的噪声信息进行标注，同时也不支持


(c) 2024 中国中文信息学会计算语言学专业委员会

蒙古语的语音情感识别任务。另一方面，尽管刘瑞 (刘瑞 et al., 2022) 等人构建了当前最大规模的蒙古语语音合成数据库，以及刘志强 (Liu et al., 2022) 等人构建了一个针对蒙古语语音识别的 IMUT-MC 语音语料库，但却没有专门针对蒙古语带噪语音情感识别领域的数据集。在方法论方面，一些工作利用语音增强或深度学习的网络架构进行带噪 SER 任务。具体而言，SR(Bandela and Kumar, 2021) 等人采用了密集的非负矩阵分解方法来对嘈杂语音信号进行去噪处理，并用支持向量机分类器对音频进行情感分类 (Jain et al., 2020)(Al Dujaili et al., 2021)。Tawari 等人 (Tawari and Trivedi, 2010) 提出了基于音调和强度轮廓的倒谱分析方法，以解决在不同环境（如汽车设置）中准确识别人类情感的挑战。MingkeXu 等人 (Xu et al., 2021) 提出了基于多头注意力机制的头部融合方法，旨在提高情感识别的准确性并增强模型在噪声环境下的鲁棒性。但在蒙古语应用场景下的带噪 SER 研究却相对滞后。这主要是因为蒙古语是一种使用人数较少的语言，因此相关的研究并不多，这导致缺乏专门针对蒙古语特性优化的情感识别模型和算法。上述问题为带噪情感识别的准确性和模型的泛化能力带来了额外的挑战。

本研究首次构建了一个专门针对蒙古语的带噪语音情感识别数据集 Noisy Mongolian Speech Emotion Recognition Dataset（NMonSER）。该数据集包含 1001 条蒙古语音频数据，覆盖 8 种情绪：生气、乏味、厌恶、高兴、自然、悲伤、害怕和惊讶，涵盖了文本、声音双模态。这些数据由一位 22 岁的专业蒙古语播音员在标准录音室内录制，并使用 Adobe Audition 软件处理，确保高质量录音。为了贴近生活中带噪环境下的音频，我们将录制的数据在不同信噪比（10db、5db、0db、-5db、-10db）下添加了 Noisex-92 标准噪声库 (Varga and Steeneken, 1993) 中的典型噪声（如 Pink 噪声、Babble 噪声等），形成了适用于噪声环境下的蒙古语音频数据集 MonSER-DN。此数据集通过精细的预处理和特征提取流程（包括音频重采样、短时傅里叶变换、Mel 频率转换等）生成了 3D 特征表示，旨在支持复杂环境下的情感识别研究。此外，文本预处理模拟了噪声环境下文本信息缺失的情况，使得数据集更贴近真实应用场景。数据集的构建考虑了真实环境中的多种挑战，如噪声干扰和信息缺失，使其成为研究蒙古语情绪识别的有价值资源。

此外，我们还提出了一种针对带噪蒙古语语音情感识别的多模态分析方法 MonSER，旨在实现带噪环境下的蒙古语语音情感识别。具体来说，我们主要通过文本特征提取、音频特征提取和模态融合三个核心模块来实现。在文本特征提取阶段，采用了 XLMBert 预训练模型 (Yuyang and Wu, 2022)，该模型通过 SentencePiece 分词和位置编码，有效捕捉蒙古语文本的语义信息。音频特征提取模块则通过卷积神经网络深入挖掘预处理音频信号的时频特征，强化了模型对音频时序变化的感知能力。最后，模态融合模块通过结合文本和音频的嵌入表示，并经过分类器进行分类，实现了高效的情感识别。整个模型不仅深入考虑了文本和语音特征的处理，还通过多个实验证明，我们的方法增强了情感识别的准确性和鲁棒性，展现了多模态情感分析在带噪蒙古语情感识别中的应用潜力。

综上，本研究不仅填补了带噪蒙古语语音情感识别领域的研究空白，而且通过开发基于语音-文本双模态的带噪语音情感识别模型 MonSER，展现了多模态情感分析在带噪蒙古语情绪识别中的应用潜力。我们的研究目标是通过自动学习最佳特征表示，有效捕获数据内部隐藏的情感信息，从而促进蒙古语技术的发展，更好地服务于少数民族人民。

本文的主要贡献如下：

1. 我们首次构建了一个面向蒙古语的单一说话人带噪语音情感识别数据集 NMonSER，填补了带噪语音情感识别领域在小语种研究方面的空白。这一数据集的构建为未来的蒙古语语音


(c) 2024 中国中文信息学会计算语言学专业委员会

情感识别研究提供了重要的资源。

2. 我们进一步提出了蒙古语带噪语音情感识别模型 MonSER，实现了带噪环境下的蒙古语情感识别，并且在模型中首次使用了 XLMBert 去提取原始蒙古语文本中的情感信息，提高了系统的鲁棒性。

3. 实验结果表明，我们的 MonSER 模型在各种噪声环境下对蒙古语语音情感识别的准确率明显高于仅使用语音作为输入的单模态语音情感识别系统，并且关于文本比例的实验也进一步验证了我们所提出方法的有效性。

## 2 数据集构建及处理

在本小节中，我们将详细介绍数据集的构建过程，包括文本收集、音频录制以及噪声添加等步骤。文本收集及音频录制等具体细节将在 2.1 小节中进行详细介绍。带噪音频制作细节等将在 2.2 小节详细介绍。

### 2.1 文本收集及音频录制

为构建 NMonSER 数据集，首先要进行文本信息的收集。选择文本的主题需要侧重于能够引发和表现出丰富情感的场景。我们从现有的蒙古文电子书、新闻以及相关影视剧中共筛选并收集了常用蒙古文 1001 条，涉及到日常生活、文化、体育、娱乐等方面的内容，这些文本中包含清晰的情感指向性，能够明确地表达或引起特定情感反应，有助于构建一个全面且效果良好的情感语料库。整个语料库数据统计情况如表 1 所示，这些文本共 (Total) 包含 82556 个字符，平均 (Mean) 每句 82 个字符，最短 (Min) 句 14 个字符，最长 (Max) 句 238 个字符；总共含有 13725 个单词，平均每句话单词量为 14 个，最大单词数量 44 个，最小单词数量 2 个。所有文本以 UTF-8 编码保存在 TXT 文件中。

| Category | Character | | | | Word | | | |
|---|---|---|---|---|---|---|---|---|
| Statistics | Total | Mean | Min | Max | Total | Mean | Min | Max |
| | 82556 | 82 | 14 | 238 | 13725 | 14 | 2 | 44 |

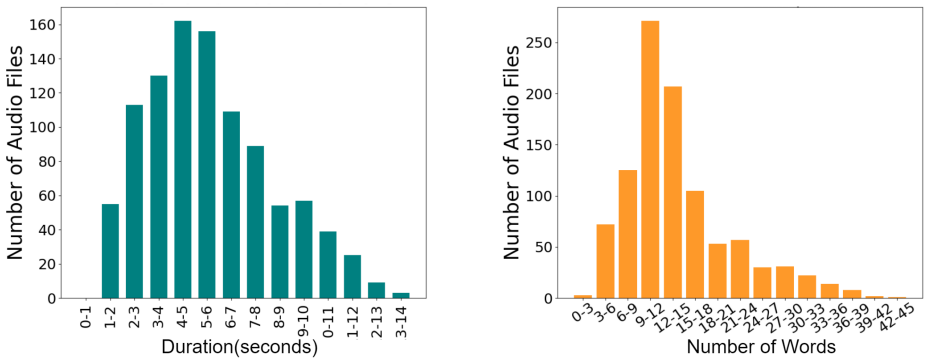
表 1: NMonSER 数据统计详情表

在语音情感识别任务中，我们避免选择过长的句子以减少情感状态的波动，确保模型更好地学习和识别情感状态。因此，如图 1 所示，我们的句子持续时间主要集中在 4 到 6 秒之间，单词数量集中在 9 到 15 之间。为了录制音频，我们邀请了一位 22 岁的蒙古语专业播音员，并有一位志愿者负责检查语法和情感表达的准确性。录音过程在标准录音工作室中进行，使用 Adobe Audition 软件进行录制。我们录制了 8 种情感的纯净音频数据，分别为生气 72 条、乏味 73 条、厌恶 71 条、高兴 72 条、自然 501 条、悲伤 72 条、害怕 72 条和惊讶 70 条，共计 1001 条蒙古语音频数据。最终录制数据总时长约 1.57 小时，所有音频以 WAV 格式存储，采样率为 44.1kHz，采样精度为 16bit。

### 2.2 噪声添加及语料库结构

为了生成不同信噪比的带噪情感语音，首先需要准备生活中的常见噪音数据和干净的情感语音数据。我们选取的噪声样本来源于 Noisex-92 标准噪声库，其中包括 Pink 噪声、Babble 噪


(c) 2024 中国中文信息学会计算语言学专业委员会




（a）MonSER-DN语料库时间长度分布直方图　（b）MonSER-DN语料库单词个数分布直方图

图 1: NMonSER 语料库数据统计图

声、M109 噪声和 F16 噪声等典型噪声样本。干净的情感音频选择上述 2.1 小节中所介绍的 1001 条蒙古语纯净的情感音频数据。其次，我们对数据集进行预处理。为了确保干净音频与噪声频率相匹配，我们将干净音频下采样至 16kHz。又由于噪声音频和干净音频的时长不能一一对应，所以我们进行了音频数据对齐操作：若干净音频数据的长度大于噪声数据长度，则利用向上取整的方法将噪声数据重复多次，直到与干净音频数据长度相匹配，确保噪声数据长度足够覆盖干净音频数据；若噪声数据的长度大于干净音频的长度，则将噪声数据截取到与干净音频数据相同的长度；若干净音频数据和噪声数据长度相同，则直接进行后续的带噪音频生成工作。接着，我们分别计算得到干净音频信号功率 $P_s$ 和噪声功率 $Pn$，具体计算如下：

$$P_s = \frac{1}{N}\sum_{n=0}^{N-1}|x(n)|^2, P_n = \frac{1}{M}\sum_{m=0}^{M-1}|y(m)|^2 \tag{1}$$

其中，$x(n)$ 表示干净音频信号，$y(m)$ 表示噪声音频信号，$N$ 是干净音频信号的长度，$M$ 是噪声音频信号的长度。然后我们通过不同信噪比（10dB、5dB、0DB、-5dB、-10dB）计算缩放系数 $k$，用于调整噪声音频的幅度，使其符合指定的信噪比，具体的计算公式如（3）：

$$k = \sqrt{\frac{Ps}{\frac{Pn}{10}(SNR/10)}} \tag{2}$$

其中，$SNR$ 是信噪比，$k$ 是用于调整噪声幅度的缩放系数，$Ps$ 是干净音频信号功率，$Pn$ 是噪声功率。最后，我们将调整好后的噪声数据与干净音频数据相加，得到混合后的音频数据。

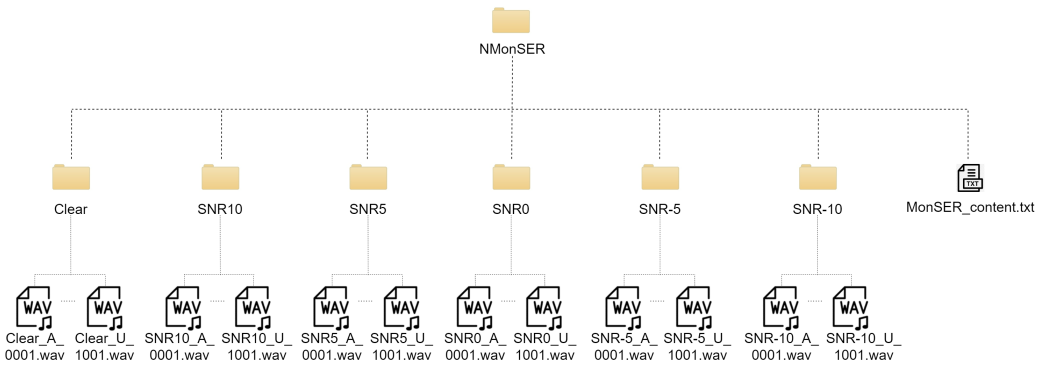



图 2: NMonSER 语料库结构图


（c）2024 中国中文信息学会计算语言学专业委员会

NMonSER 语料库的文件结构如图 2 所示，将叠加噪声后的音频数据保存为 WAV 格式的文件，采样率为 16kHz，编码为 16 位。为了更好地管理数据，我们根据不同的信噪比对这些文件进行分类，并将它们放置在相应命名的文件夹中，每个文件夹中有 1001 条音频。此外，我们使用了 ID 序列来命名文本语料，并将其存储在名为 MonSER_content 的 txt 文件中。对于音频文件的命名，则是由信噪比、情绪标签以及文本 ID 序列组合而成。这样的文件管理结构有助于我们更加清晰地组织和管理 NMonSER 语料库中的数据。

## 3 模型架构

在本节中，提出并详细解释了所提出的 MonSER 模型，该模型的框架结构如图 3 所示。主要由三个模块所构成，包含蒙古文文本编码器、蒙古语语音编码器和语义增强模块。

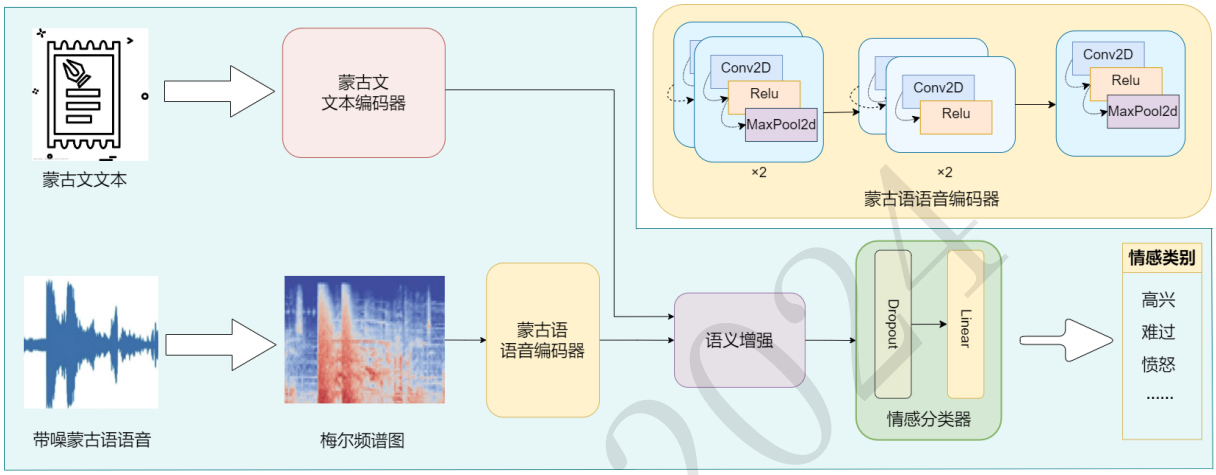



图 3: MonSER 模型结构图

### 3.1 蒙古文文本编码器

在蒙古文文本编码器中，我们主要运用了 XLMBert 对文本进行处理。XLMBert 是基于 BERT 模型的预训练语言模型，专门用于处理多语言文本，包括蒙古语。它通过在大规模文本语料上进行预训练来学习通用的语言表示，然后可以在各种下游任务上进行微调。XLMBert 对蒙古语的编码过程，大致包括以下步骤：首先，XLMBert 使用 SentencePiece 分词器对蒙古语文本进行分词处理，将文本切分为词片段（subword units）。然后，将分词后的词片段映射为词嵌入向量。XLMBert 模型在预训练过程中学习了词嵌入向量，能够将每个词片段转换为一个固定维度的向量表示，包含了词汇信息和语义信息。接下来，在输入序列中加入位置编码，以区分不同位置的词片段。位置编码通常是一种特定模式的向量，用于表示每个词片段在序列中的位置信息。将词嵌入向量与位置编码相加，得到最终的输入序列表示。这个输入序列会经过多层 Transformer 编码器，进行多层的自注意力机制和前向传播操作，从而逐步提取输入文本的语义特征，该语义特征最终会被表示为 768 维向量。

### 3.2 蒙古语语音编码器

蒙古语语音编码器使用 3D 频谱图作为输入。相较于 2D 特征，3D 特征在处理时序数据和捕捉时序信息方面具有优势 (Krizhevsky et al., 2012)，能有效捕获数据在时间维度上的演变和


(c) 2024 中国中文信息学会计算语言学专业委员会

变化，对蒙古语音情感识别至关重要。模块主要定义了卷积神经网络模型，接收经过短时傅里叶变换 STFT 处理的频谱图，通过一系列特征提取层逐步加工 3D 频谱图特征。

具体来说，首层卷积使用了 64 个 11x11 滤波器，步长为 4，填充为 2，旨在捕获 3D 频谱图的低级特征，如边缘和纹理。ReLU 激活引入非线性，最大池化层减小特征图尺寸，保留关键信息。随后，通过更细致的卷积进一步提炼特征，包括第二层 192 个 5x5 滤波器，提高对高级模式的识别能力。第三至五层卷积持续深化对 3D 频谱图的理解，包括第三个卷积层 384 个 3x3 滤波器，以及第四和第五个卷积层各 256 个 3x3 滤波器，旨在捕获更为复杂和抽象的特征，为最终任务提供丰富的特征集。最后的池化层确保特征图的尺寸进一步减小，同时保留最为重要的特征信息。整个特征提取流程有效地从 3D 频谱图中提炼出关键特征，形成 1*256 的数组，作为下一步模态融合的输入条件。通过对每一层的参数和结构优化，模型能够根据具体任务和数据集的需求达到最佳性能。

### 3.3 语义增强模块

在语义增强模块，主要目标是实现文本和音频信息的有效融合，并对这些融合后的信息进行神经网络处理 (Christy et al., 2020)。首先，将经过编码的文本和音频嵌入表示沿列方向拼接，生成一个综合的特征表示。这种融合方式旨在创建一个全面丰富的特征空间，为后续的情感分类任务奠定基础。接下来，通过 Dropout 层对这个融合的特征表示进行随机的失活处理，以此来减轻模型的过拟合现象。Dropout 通过随机使一部分神经元的输出为零，降低了模型对单一数据特征的依赖，增强了模型的泛化能力。最后，linear 线性层将经 Dropout 处理的特征数据映射到一个新的空间，用于输出最终的分类结果。这一步是通过一个全连接的线性层完成的，它将融合后的特征转换成模型预测所需的具体输出形式，确保了模型可以对不同的信息源进行有效的学习和整合。

综上所述，本节主要介绍了蒙古语语音情感识别设计的 MonSER 模型，包括蒙古文文本编码器、蒙古语语音编码器和语义增强模块。蒙古文文本编码器主要用 XLMBert 模型处理，利用 SentencePiece 分词器进行分词、映射词嵌入向量、添加位置编码，最终提取深层语义特征。蒙古语语音编码器主要利用卷积层等结构提取频谱图特征，捕获语音情感特性。语义增强模块将文本和音频信息结合，形成综合特征，通过 Dropout 层减少过拟合，线性层映射为情感分类输出。这一设计有效地实现了噪声环境下对蒙古语语音情感的识别。

## 4 实验设计

### 4.1 数据预处理

数据预处理主要包括对原始音频的预处理和蒙古语文本的预处理。首先，对原始音频进行标准化处理，生成对应的频谱图以及其对应的 3D 特征表示。其次，对蒙古语文本进行处理，在实际环境中，由于噪声的存在，获取的文本信息可能会有一定程度的缺失，因此我们分别以不同比例和不同单词个数将文本中的单词替换为'[mask]'，从而模拟文本信息缺失的情况。

#### 4.1.1 音频预处理

对原始音频进行预处理时，首先使用 Python 中 SciPy 库中的 wavfile.read 函数，获取音频文件的采样率和音频数据，对其进行重采样、分帧等预处理。然后对每一帧的音频信号进行短时


(c) 2024 中国中文信息学会计算语言学专业委员会

傅里叶变换（STFT），得到其频域信息，其离散形式的计算公式为：

$$X(m,\omega) = \sum_{n=0}^{N-1} x(n)w(n-m)e^{-j\omega n} \tag{3}$$

其中，$X(m,\omega)$ 表示频率 $\omega$ 处在时间片段 $m$ 的频谱值，$x(n)$ 是输入信号的离散时间序列，$w(n-m)$ 是窗口函数，$N$ 是窗口长度。

然后使用 Mel 滤波器组对频域信息进行加权，得到每个 Mel 频率段的能量，计算公式为：

$$M(f) = 2595\log_{10}(1+\frac{f}{700}) \tag{4}$$

其中，$M(f)$ 表示频率 $f$ 对应的 Mel 频率。

接着再将每个 Mel 频率段的能量取对数，得到 Log-Mel 频谱图，计算公式为：

$$log-Mel(f) = \log(1+\frac{E(f)}{E_0}) \tag{5}$$

其中，$E(f)$ 表示 Mel 频率段 $f$ 的能量，$E_0$ 是一个小常数，用于防止对数变换时出现零值。

最后，为了更好的处理时序数据和捕捉时序信息，我们根据提供或默认的均值和标准差，对输入的频谱图进行标准化处理，将得到的语谱图转换为 3D 特征表示，用于输入模型中进行训练和预测。

#### 4.1.2 文本预处理

在实际应用场景中，面对环境噪音等因素的干扰，我们常常会遇到文本信息不完整的挑战(Fan et al., 2023)(Liu et al., 2024)。为了模拟这种情况并深入理解在含噪环境下处理文本的复杂性，我们采纳了一项独特的方法。具体来说，我们通过将文本中的文字以不同的比例（20%、40%、60%、80%）替换为'[mask]' 标记，有效地重现了文本信息丢失的情境。具体操作如图 4(a) 所示。

然而，这种方法在处理少于 5 个词的短句时遇到了困难，主要表现在对短句进行'[mask]' 替换时，不同比例下可能出现相似或相同的情况。例如，对于仅含有四个词的句子，当替换比例为 20% 时，可能会与原句完全相同；对于三个词的句子，20% 的替换比例下也可能出现相同的现象，而 40% 和 60% 的替换比例则呈现相似的替换效果；最极端的情况是，对于只有两个词的句子，在 20% 和 40% 的替换比例下可能完全不变，而在 60% 和 80% 的替换比例下也可能出现相似的替换效果。这种现象可能导致 XLMBert 在处理不同文本缺失比例的情况下生成相似或相同的编码，进而在含噪声环境下的语音情感识别中得到类似的准确率，影响对模型性能的评估。为了应对这一挑战，我们进一步筛选了长度超过 5 个词的数据样本，并采用更细致的掩码策略，即在每个句子中随机替换 1 到 5 个词，具体操作如图 4(b) 所示。通过这种方法生成的新数据集，我们能够更准确地模拟和训练模型，以提升结果的可信度。

### 4.2 参数配置

在本实验中，我们采用了 XLMBert 对文本数据进行处理和编码，提取的文本特征为 768 维。音频特征提取部分采用了一系列卷积神经网络层和最大池化层，从输入的音频数据中提取特征，音频特征维度为 256 维。接着，我们将融合后的文本和音频特征映射到 1024 维度的特征空间，通过全连接层进行分类，得到最终的输出结果。具体模型参数如下：首先是卷积层，输入通道为 3，输出通道为 64，卷积核大小为 11x11，步长为 4，填充为 2，采用 ReLU 激活函数。


(c) 2024 中国中文信息学会计算语言学专业委员会

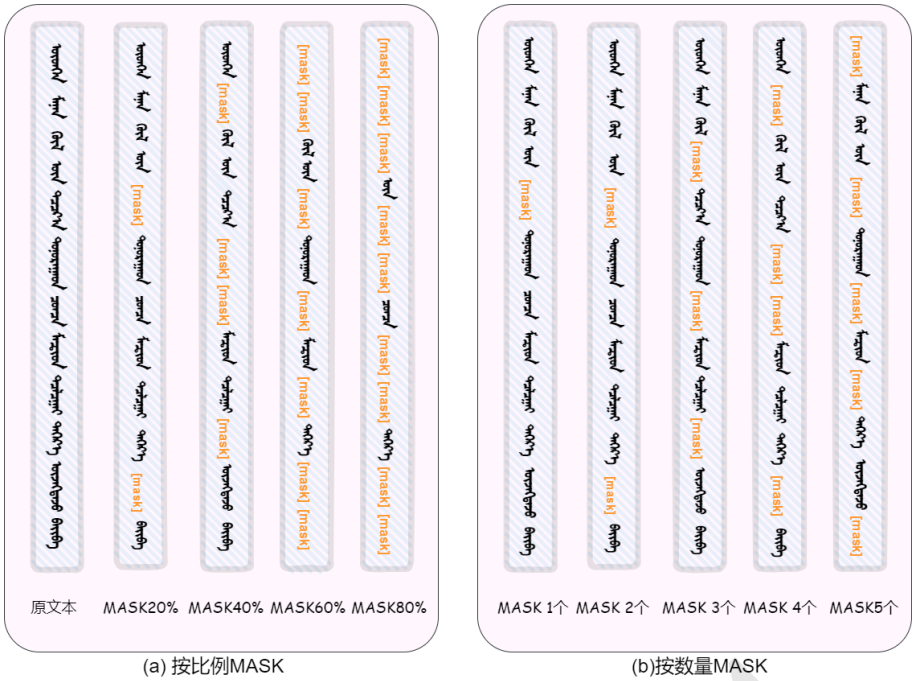



图 4: 蒙古文文本 mask 策略，(a) 按比例 MASK (b) 按数量 MASK

接着是最大池化层，池化核大小为 3x3，步长为 2。之后的卷积层和 ReLU 激活函数与前述类似，用于进一步提取和加工音频特征。紧接着是自适应平均池化层，将特征图大小调整为 12x12。然后是全连接线性层，将融合后的文本和音频特征映射到 1024 维度的特征空间。为了防止过拟合，我们使用了 Dropout 层，在训练过程中随机失活一部分神经元。最后是全连接线性层，用于最终的分类任务输出，输入特征维度为 1280。我们使用了学习率为 0.001 的 Adam 优化器，并采用了交叉熵损失函数。Batch size 设置为 32，Dropout 率为 0.5。初始的 epoch 值设为 150，但我们实际上应用了提前停止（early stopping）函数，在损失连续 20 次未下降时停止模型训练。

### 4.3 实验结果

实验结果如表 2 所示。第一行展示了干净语音以及不同信噪比下的带噪音频，信噪比分别为 10db、5db、0db、-5db、-10db。第二行表示仅包含语音信息时对应各类带噪音频的语音情感识别准确度。从表格中可以明显看出，随着信噪比逐渐降低，即噪声逐渐增大，情感识别的准确率也显著下降，从干净语音的 0.9108 降至信噪比为-10db 时的 0.6832。

| SNR / MASK | clear | 10db | 5db | 0db | -5db | -10db |
|---|---|---|---|---|---|---|
| - | 0.9108 | 0.8989 | 0.7921 | 0.7822 | 0.7426 | 0.6832 |
| 0% | 0.9900 | 0.9703 | 0.9307 | 0.9109 | 0.8400 | 0.7822 |
| 20% | 0.9800 | 0.9600 | 0.9208 | 0.9010 | 0.8300 | 0.7600 |
| 40% | 0.9505 | 0.9307 | 0.9208 | 0.8911 | 0.8218 | 0.7524 |
| 60% | 0.9200 | 0.9109 | 0.8700 | 0.8416 | 0.8119 | 0.7200 |
| 80% | 0.9192 | 0.9091 | 0.8317 | 0.8300 | 0.8020 | 0.7129 |

表 2: 蒙古文按比例 MASK 与不同信噪比音频上的情感识别准确率


(c) 2024 中国中文信息学会计算语言学专业委员会

表格的第一列显示了无文本信息以及在噪声环境下丢失的文字信息比例，分别达到 0%、20%、40%、60%、80%。加入文本信息后，各类噪声环境下的准确率明显提升，精度提高了 7.14% 到 13.86%。这表明即使在文本信息有所缺失的情况下，各类噪声环境下的情感识别精度仍然高于单一模态下的准确率。然而，在信噪比为 5db 时，当文字缺失比例达到 20% 和 40% 时，对应的精度是相同的。这种现象可以归因于之前提到的短句问题。当句子过短时，对句子按不同比例进行 mask 可能导致 mask20% 和 mask40% 所对应的文本有效信息相近，这可能导致在同一信噪比下，不同 mask 率得到的准确率相同的情况发生。

为了解决上述现象对模型评估的影响，我们采取了一项重要措施：剔除长度小于等于 5 的句子，并仅保留了长度大于 5 的 959 条有效数据，随后再次进行分类实验。具体的结果如表 3 所示。在这个表格中，第一行与之前的表格相同，第一列表示在一句话中丢失的单词数量，范围从 1 个至 5 个递增。

| SNR / MASK WORDS | clear | 10db | 5db | 0db | -5db | -10db |
|---|---|---|---|---|---|---|
| 0 | 0.9900 | 0.9792 | 0.9583 | 0.9375 | 0.8646 | 0.8021 |
| 1 | 0.9896 | 0.9583 | 0.9479 | 0.9271 | 0.8438 | 0.7917 |
| 2 | 0.9688 | 0.9479 | 0.9271 | 0.9062 | 0.8333 | 0.7708 |
| 3 | 0.9583 | 0.9375 | 0.9167 | 0.8958 | 0.8125 | 0.7500 |
| 4 | 0.9375 | 0.9271 | 0.9062 | 0.8542 | 0.7917 | 0.7188 |
| 5 | 0.9271 | 0.9167 | 0.8958 | 0.8333 | 0.7604 | 0.7083 |

表 3: 蒙古文按单词个数 MASK 与不同信噪比音频上的情感识别准确率

通过按照缺失单词个数进行分类实验，我们观察到在同一信噪比下，不同 mask 率得到的准确率不再相同。这说明我们的模型在噪声环境下进行蒙古语语音情感识别的鲁棒性较好，并且具有较高的精度。这一实验结果进一步验证了我们所提出的模型在处理噪声环境下的可靠性和稳健性。

## 5 结论

综上所述，我们的基于文本-语音双模态的带噪蒙古语语音情感识别模型在实验中取得了较为不错的结果。通过构建数据集、优化特征提取和融合策略，我们成功提高了蒙古语情感识别的准确性和鲁棒性。具体来说，我们通过积极收集和整理蒙古语情感数据，构建了一个适用于情感识别研究的重要资源，为该领域的进一步发展提供了基础。在模型设计方面，我们充分利用了文本信息和语音特征的双模态优势，通过 XLMBert 模型对蒙古语文本进行编码，并结合音频特征提取器对频谱图进行分析，从而实现了对语音情感的精准识别。这项工作不仅在技术上取得了重要进展，也为蒙古语情感识别领域的研究和实际应用带来了新的启示。我们的成功经验和成果将为未来相关工作提供有益的借鉴和参考，推动语音情感识别技术在蒙古语言言等小语种领域的不断创新和应用。

第二十三届中国计算语言学大会论文集，第268页-第279页，太原，中国，2024年7月25日至28日。
卷1：主会论文
(c) 2024 中国中文信息学会计算语言学专业委员会

## 致谢

感谢国家自然科学基金青年科学基金项目（N0.62206136）；广东省数字孪生人重点实验室（华南理工大学）开放课题（2022B1212010004）；“一区两基地”超算能力建设项目（内蒙古大学）（项目编号：21300-231510）的资助。

## 参考文献

Chenchah, Farah, and Zied Lachiri. 2016. Speech emotion recognition in noisy environment. In *Proceedings of the 2016 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*. IEEE.

Wei Xu. 2019. Toward human-centered AI: a perspective from human-computer interaction. *interactions*, 26(4):42–46.

Wei Bi, et al. 2022. Enterprise strategic management from the perspective of business ecosystem construction based on multimodal emotion recognition. *Frontiers in Psychology*, 13:857891.

Parian Haghighat, et al. 2023. Effects of an intelligent virtual assistant on office task performance and workload in a noisy environment. *Applied Ergonomics*, 109:103969.

Maria Lucia Barron-Estrada, Ramon Zatarain-Cabada, and Raul Oramas Bustillos. 2019. Emotion Recognition for Education using Sentiment Analysis. *Res. Comput. Sci.*, 148(5):71-80.

Weiqing Wang, et al. 2020. Emotion recognition of students based on facial expressions in online education based on the perspective of computer simulation. *Complexity*, 2020:1-9.

Oussama El Hammoumi, et al. 2018. Emotion recognition in e-learning systems. In *Proceedings of the 2018 6th International Conference on Multimedia Computing and Systems (ICMCS)*, IEEE.

Sonali Singh and Navita Srivastava. 2023. Emotion recognition for mental health prediction using AI techniques: an overview. *International Journal of Advanced Research in Computer Science*, 14(3).

Daniel Cen Ambrus. 2000. Collecting and recording of an emotional speech database. Maribor, Slovenia: University of Maribor.

Sinéad McGilloway, et al. 2000. *Approaching automatic recognition of emotion from voice: A rough benchmark*. ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion.

Ellen Douglas-Cowie, Roddy Cowie, and Marc Schrö der. 2000. *A new emotion database: considerations, sources and scope*. ISCA tutorial and research workshop (ITRW) on speech and emotion.

Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Wolfgang Sendlmeier, and Benjamin Weiss. 2005. A database of German emotional speech. In *Interspeech*, Vol. 5.

Cheng-Lin Liu, Fei Yin, and Dong-Yan Huang. 2011. CASIA online and offline Chinese handwriting databases. In *2011 International Conference on Document Analysis and Recognition*, IEEE.

Carlos Busso, Murtaza Bulut, et al. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335-359.

刘瑞, 康世胤, 李劲东, 飞龙, 高光来. 2022. MonTTS: 完全非自回归的实时, 高保真蒙古语语音合成模型. 中文信息学报, 36(7):86–97.

赵建东, 高光来, 飞龙. 2014. 基于 HMM 的蒙古语语音合成技术研究. 计算机科学, 41(1):80–82.

Zhiqiang Liu, et al. 2022. IMUT-MC: A speech corpus for Mongolian speech recognition. *China Sci. Data*, 7:13.

Surekha Reddy Bandela and T. Kishore Kumar. 2021. Unsupervised feature selection and NMF denoising for robust Speech Emotion Recognition. *Applied Acoustics*, 172:107645.

Ashish Tawari and Mohan Manubhai Trivedi. 2010. Speech emotion analysis: Exploring the role of context. *IEEE Transactions on Multimedia*, 12(6):502-509.


(c) 2024 中国中文信息学会计算语言学专业委员会

Mingke Xu, Fan Zhang, and Wei Zhang. 2021. Head fusion: Improving the accuracy and robustness of speech emotion recognition on the IEMOCAP and RAVDESS dataset. *IEEE Access*, 9:74539-74549.

Andrew Varga and Herman JM Steeneken. 1993. *Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems.* Speech communication, 12(3): 247-251.

Deng Yuyang and Dan Wu. 2022. *Chinese-Tibetan Bilingual Named Entity Recognition for Traditional Tibetan Festivals.* Data Analysis and Knowledge Discovery, 7(7): 125-135.

A. Christy, et al. 2020. *Multimodal speech emotion recognition and classification using convolutional neural network techniques.* International Journal of Speech Technology, 23(2), 381-388.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. *Imagenet classification with deep convolutional neural networks.* Advances in neural information processing systems 25.

Youngja Nam and Chankyu Lee. 2021. *Cascaded convolutional neural network architecture for speech emotion recognition in noisy conditions.* Sensors, 21(13):4399.

Upasana Tiwari, et al. 2020. *Multi-conditioning and data augmentation using generative noise model for speech emotion recognition in noisy conditions.* In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE.

Htwe Pa Pa Win and Phyo Thu Thu Khine. 2020. *Emotion recognition system of noisy speech in real world environment.* International Journal of Image, Graphics and Signal Processing (IJIGSP), 12(2):1-8.

Abeer Ali Alnuaim, et al. 2022. *Human-computer interaction for recognizing speech emotions using multilayer perceptron classifier.* Journal of Healthcare Engineering, 2022.

Qi Fan, et al. 2023. *Learning Noise-Robust Joint Representation for Multimodal Emotion Recognition under Realistic Incomplete Data Scenarios.* arXiv preprint arXiv:2311.16114.

Rui Liu, et al. 2024. *Contrastive Learning based Modality-Invariant Feature Acquisition for Robust Multimodal Emotion Recognition with Missing Modalities.* IEEE Transactions on Affective Computing.

Ming-Che Lee, et al. 2020. *Study on emotion recognition and companion Chatbot using deep neural network.* Multimedia Tools and Applications, 79(27), 19629-19657.

Rajdeep Chatterjee, et al. 2021. *Real-time speech emotion analysis for smart home assistants.* IEEE Transactions on Consumer Electronics, 67(1), 68-76.

Sharif Noor Zisad, Mohammad Shahadat Hossain, and Karl Andersson. 2020. *Speech emotion recognition in neurological disorders using convolutional neural network.* In International Conference on Brain Informatics, Cham: Springer International Publishing.

Manas Jain, et al. 2020. *Speech emotion recognition using support vector machine.* arXiv preprint arXiv:2002.07590.

Mohammed Jawad Al Dujaili, Abbas Ebrahimi-Moghadam, and Ahmed Fatlawi. 2021. *Speech emotion recognition based on SVM and KNN classifications fusion.* International Journal of Electrical and Computer Engineering, 11(2), 1259.

CCL2024

# 基于神经编解码语言模型的老挝语韵律建模方法

**易宁静[1,2]，王琳钦[1,2]，高盛祥*[1,2]，余正涛[1,2]**
1.昆明理工大学，信息工程与自动化学院，昆明，650500
2.昆明理工大学，云南省人工智能重点实验室，昆明，650500
897079505@qq.com,2424172505@qq.com,gaoshengxiang.yn@foxmail.com,
ztyu@hotmail.com

## 摘要

为了赋予合成语音类似人类语言的丰富韵律和节奏变化，现有方法普遍采用基于随机数的时长预测器。这些方法通过使用随机数初始化的潜在变量来模拟人类说话的多样节奏变化。然而，由于依赖于随机数噪声的局限性，这些方法合成的语音往往仍然缺乏真实语音的多样性和韵律变化的丰富性。与之前方法不同，本文提出了一种基于神经编解码语言模型（VALL-E）的韵律建模方法，本文利用先验速度和音调时序变化曲线建模韵律变化分布，有效融入神经编解码语言模型训练过程中，并且在推理阶段可通过控制先验时序曲线控制生成语音的韵律。实验证明，本文方法合成英语音频达到了4.05的MOS评分，合成老挝语音频达到了3.61的MOS评分。基于神经编解码语言模型的老挝语韵律建模方法，能很好的在速度和音调方面实现韵律的可控性。

**关键词：** 韵律；老挝语；神经编解码语言模型；可控性

# A Method for Lao Prosody Modeling Based on Neural Codec Language Model

**Ningjing Yi[1,2], Linqin Wang[1,2], Shengxiang Gao*[1,2], Zhengtao Yu[1,2]**
1.Faculty of Information Engineering and Automation,
Kunming University of Science and Technology, Kunming 650500, China
2.Yunnan Key Laboratory of Artificial Intelligence,
Kunming University of Science and Technology, Kunming 650500, China
897079505@qq.com,2424172505@qq.com,gaoshengxiang.yn@foxmail.coma,
ztyu@hotmail.com

## Abstract

In order to give synthesized speech rich prosody and rhythm changes similar to human language, existing methods generally use duration predictors based on random numbers. These methods simulate the diverse rhythmic variations of human speech by using latent variables initialized with random numbers. However, due to the limitations of relying on random number noise, the speech synthesized by these methods often still lacks the diversity of real speech and the richness of prosodic changes. Different from previous methods, this paper proposes a prosody modeling method based on the neural codec language model (VALL-E). This paper uses a priori speech speed and pitch timing change curves to model the distribution of prosody changes, effectively integrating into the neural codec language During the model training process

*高盛祥（通信作者）：gaoshengxiang.yn@foxmail.com
**基金项目**：国家自然科学基金（62376111，U23A20388，U21B2027，62366027）；云南省重点研发计划（202303AP140008，202302AD080003，202401BC070021，202103AA080015）；云南省科技人才与平台计划（202105AC160018）
©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

and during the inference stage, the prosody of the generated speech can be controlled by controlling the prior timing curve. Experiments have shown that the method in this article achieves a MOS score of 4.05 when synthesizing English audio, and a MOS score of 3.61 when synthesizing Lao audio. The Lao prosody modeling method based on the neural encoding and decoding language model can achieve good controllability of prosody in terms of speaking speed and pitch.

**Keywords:** Prosody , Lao , Neural Codec Language Model , Controllability

## 1 引言

老挝与中国比邻，2021年12月3日中老铁路的开通进一步加深了中国与老挝的国际交流，开展老挝语的韵律建模研究对促进两国政策沟通有重要支撑，符合国家“一带一路”建设等战略需求。近年来，语音合成模型在合成声音的自然度方面有了极大的提高，跳词复读的问题也得到了解决。因此，如何在合成语音中实现人类语音的表现力是一个研究热点，而让语音富有表现力的关键，就是对语音进行韵律建模。

相较于中、英等语言，老挝语与它们的发音存在一定的区别。老挝语的字符读法分为元音和辅音两类，元音分长短两类，共28个，辅音共27个。另一方面，老挝语是音调语言，音调会直接添加到字符的上方，音调的改变会改变词语本身的意思，这使得老挝语的语音合成需要在音节及音调上准确建模。最近，(Anh and Thanh, 2022)首次实现了基于神经网络的老挝语语音合成，但这项研究工作只是在基准模型上实现了老挝语的语音合成，在合成语音的韵律方面并没有达到很好的效果。因此，针对老挝语的韵律建模仍是一个值得探索的问题。

韵律是语言中多种特征的融合，韵律建模内容可包括速度、音高、持续时间等等信息，主要使合成的语音更加拟人化。对于韵律建模任务，按照韵律调整的粒度可分为两类：(1)粗粒度：粗粒度为句子层面的迁移调控。(2)细粒度：细粒度为短语，单词层面的调控。在粗粒度方面，(Skerry-Ryan et al., 2018)等人提出的参考韵律嵌入方法，首次使用无监督学习来进行韵律的迁移，把参考的语音编码成一个向量，向量包含参考语音的韵律信息。(Wang et al., 2018)等人提出的风格token可以对韵律信息进行解耦，使其每个token控制一种风格。但存在韵律信息无法完全解耦，每个token仍包含多种信息的问题。在细粒度方面，(Lee and Kim, 2019)等人在帧级别和音素级别，使用参考编码器进行细粒度的韵律建模。但缺点是对未见说话人进行韵律迁移效果较差。(Klimkov et al., 2019)提出了对参考音频进行单独的信息抽取方法，并通过变分自编码器模型(VAE)对韵律信息进行预测，改进了对未见说话人的韵律迁移效果。这些工作在韵律迁移方面已经取得了较好的效果，但在韵律的可控性方面并没有受到太多关注。在语言交流中，韵律的可控性对于传达说话者的情感、意图甚至语义至关重要。例如，放慢语速和提升音调有助于突出某一个词语，强调这个词语的重要性。因此，实现对韵律的可控性能使语音合成技术在许多应用场景中发挥更好的效果，这正是本文所关注的。

为了解决老挝语韵律建模任务中对语言特征缺乏可控性等问题，受ControlVC(Chen and Duan, 2022)的启发，本文提出了一种可以对速度和音调进行时变控制的老挝语语音合成模型。本文在神经编解码语言模型VALL-E(Wang et al., 2023)的架构体系上，利用先验速度和音调时序变化曲线建模韵律变化分布，随后分别在英语和老挝语数据集上进行训练，最终提出韵律可控的语音合成模型。本文的贡献如下：

(1)提出了韵律可控的语音合成方法，实现了在速度和音调多种语言特点上的可控性韵律建模，解决了语音合成任务中合成语音韵律不丰富、不可控的问题。

(2)构建不同的字符转音素方法，利用大规模多语言、多后端字符转音素系统Epitran构建老挝语音素字典，实现了在神经编解码语言模型上的低资源语言老挝语的韵律建模。

(3)在585小时的英语数据集上的合成语音达到4.05的MOS评分，在50小时左右的老挝语数据集上的合成语音达到3.61的MOS评分，

## 2 相关工作

(1)传统的端到端(End-to-End)语音合成模型通常由文本分析、声学模型、声码器三个部分组成，基于端到端语音合成模型的韵律建模通常是在声学模型部分进行修改。有通过全局风格


(c) 2024 中国中文信息学会计算语言学专业委员会

标记(Wang et al., 2018)的方法，训练不需要真实标签，模型学习将各种噪声和说话者因素分解为单独的风格标记；有通过韵律调控(Lee and Kim, 2019)的方法，即在嵌入网络中引入时间结构，从而实现对合成语音说话风格的细粒度建模；有通过隐变量(Kim et al., 2021)而非频谱串联起来声学模型和声码器的方法，在隐变量上进行随机建模并利用随机时长预测器，对于同样的输入文本，能够合成不同韵律的语音，提高了合成语音的多样性。但由于其依赖于随机数噪声的局限性，合成的语音仍然缺乏真实语音的多样性和韵律变化的丰富性。ControlVC(Chen and Duan, 2022)提出了一种对音调和速度进行时变控制的神经语音转换模型。ControlVC使用控制曲线对音调和速度进行时变控制，使用预先训练的编码器来生成音调和语言嵌入，使用说话人编码器生成说话人嵌入，最后使用声码器将其组合并转换为目标语音。但其训练数据量小，ControlVC使用的公共英语数据集仅包含约44小时的英语语音。且上下文学习能力弱，并不能在推理过程中转换出未见说话人的语音。

端到端语音合成模型的优点是简化了语音合成系统中的流程，减少了多个阶段的处理步骤，使得系统更加简洁高效。但该方法合成语音的质量还有待提升。

(2)神经编解码语言模型(Neural Codec Language Model)引入了一种用于语音合成的语言建模方法，将语音合成视为条件语言建模任务。与端到端语音合成模型相比，神经编解码语言模型的中间表示不再是梅尔谱图，而是离散的编码。具体来说，神经编解码语言模型引入了一种基于提示的语音合成方法，以EnCodec(Défossez et al., 2022)的音频编解码器作为中间表示。通过将语音波形编码为离散声学代码，这些模型能够重建高质量的波形，即使对于未见说话人也是如此。基于神经编解码语言模型的韵律建模也已经开展了一些工作，例如SC VALL-E(Kim et al., 2023)，它从文本句子和提示音频中获取输入，不再是简单地模仿提示音频的特征，而是通过控制属性来产生不同的声音。

神经编解码语言模型的优点是训练数据量大，且上下文学习能力强。对于语音合成任务而言，如果模型无需微调即可为训练中未出现过的说话人合成高质量的语音，则该模型被认为具有上下文学习能力。而神经编解码语言模型无需微调，只需输入未见说话人的3秒语音，就能合成出高质量的音频。

## 3 方法

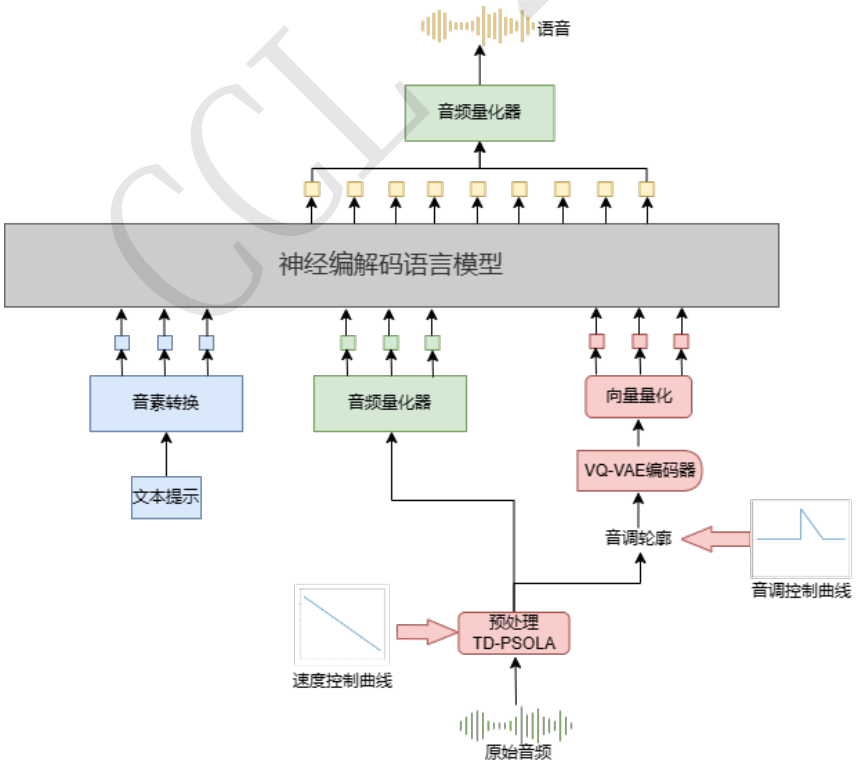



图 1: 模型总图


(c) 2024 中国中文信息学会计算语言学专业委员会

本文旨在利用控制曲线在英语和老挝语语音合成中实现韵律的时变控制。本文基于VALL-E的模型结构，受ControlVC(Chen and Duan, 2022)的启发，如图1所示，系统由三个部分组成：速度预处理和音素转换部分、音频量化和音调嵌入部分、音频合成部分。在速度预处理和音素转换部分，采用TD-PSOLA算法，根据速度控制曲线修改输入语音的语速，并且将输入的文本转换为音素。在音频量化和音调嵌入部分，将预处理后的音频分别送入神经音频编解码器和VQ-VAE编码器，分别转换为离散声学代码及离散音调嵌入。最后，在音频合成部分，将离散代码重构为语音波形。

### 3.1 速度预处理和音素转换

在预处理阶段，我们使用时域音调同步重叠相加（TD-PSOLA）算法(Charpentier and Stella, 1986)，根据输入的速度控制曲线修改语音的语速。我们首先对输入的音频重采样至24kHz，再对音频进行分段，并使用控制曲线相应位置指示的拉伸比率对每一段音频的时间进行拉伸或收缩。遵循ControlVC，如图2所示，速度控制曲线的三种模式，分别为上升、下降和抛物线。具体来说，首先我们对原始音频进行分段，再使用np.array函数创建整数序列，并根据线性公式或抛物线公式计算每一段的速度，使语音的速度可以随着时间越来越快，越来越慢，也可像抛物线一样先快后慢。在这一阶段，语音的音色、音调均被保留。

基准模型VALL-E使用的字符转音素方法并不支持老挝语，因此，我们构建特殊的字符转音素模块来进行老挝文本的字符转音素。在这一模块中，我们引入了Epitran(Mortensen et al., 2018)，Epitran是一个用于字符转音素的大规模多语言、多后端系统，支持包括老挝语在内的多种语言，它采用语言正字法中的单词标记并输出音素表示，我们再将其输出的音素表示构成老挝语音素词典。最终达到捕捉老挝语语言的差异性，模型更好地学习老挝语语言知识的效果。

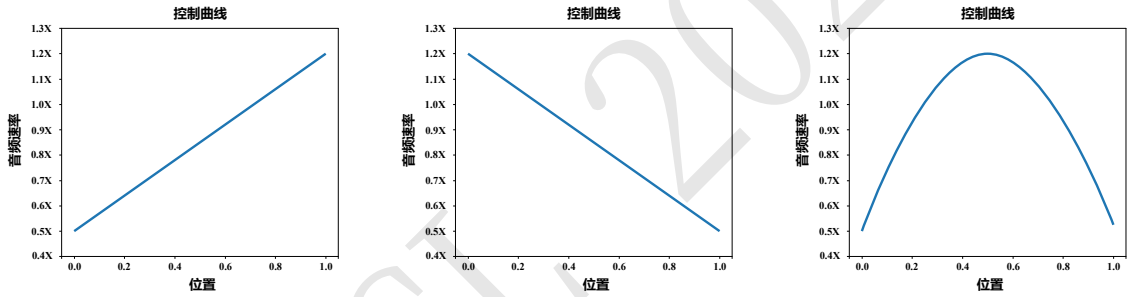



图 2: 速度控制曲线

### 3.2 音频量化和音调嵌入

由于音频通常存储为16位整数值序列，因此需要生成模型在每个时间步输出$2^{16} = 65536$个概率来合成原始音频。此外，超过万的音频采样率导致序列长度过长，使得原始音频合成变得更加困难。为此，需要音频量化来压缩整数值和序列长度。在本文中，我们遵循AudioLM(Borsos et al., 2023)，利用神经编解码器模型来表示离散标记中的音频。为了压缩音频以进行网络传输，编解码器模型能够将输入的语音波形编码为离散声学代码，对于训练中未出现的说话者，也能重建出高质量波形。神经编解码器可以减少时间步长以提高效率，并且它包含丰富的说话人信息和声学信息。在本文中，我们采用预先训练的神经音频编解码器模型EnCodec(Défossez et al., 2022)作为我们的量化器。EnCodec是一种卷积编码器-解码器模型，其输入和输出都是可变比特率的采样率为24kHz的音频。编码器将24kHz的输入波形产生75Hz的嵌入，采样率降低了320倍。其中量化器共有八层，每个嵌入均通过残差向量量化(RVQ)进行建模。

我们采用YAAPT算法(Kasi and Zahorian, 2002)提取经过速度控制的语音的音调序列$(p_1, ..., p_T)$，其中T 是帧。然后将该音调序列乘以输入的音调控制曲线以获得修改的音调序列$(p'_1, ..., p'_T)$。同样遵循ControlVC，输入的音调控制曲线示例如图3所示，模式为stress模式，该曲线表示修改后的语音与原始语音的基频(F0s，单位为Hz)的比率。例如，1.2x F0率表


(c) 2024 中国中文信息学会计算语言学专业委员会

示修改后的语音的F0预计为原语音的1.2倍。这里的实现思路是，在音调序列中，选取一段区间作为上升的部分，对该部分音调进行线性调整，使得音调值随着位置的增加而线性上升。修改后的音调序列作为基于VQ-VAE编码器的音调嵌入网络(Dhariwal et al., 2020)的输入，以获得修改后的音调嵌入。编码器产生潜在向量序列$(h_1, ..., h_T)$，并将其映射到码本中各自最接近的代码。然后，我们采用码本向量的整数索引来形成音调嵌入序列$z^{(p)} = (z_1^{(p)}, ..., z_T^{(p)})$。

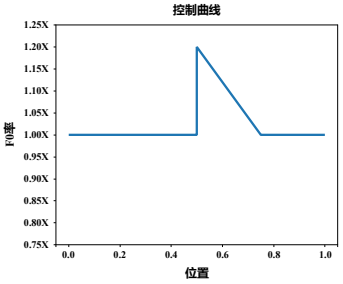



图 3: 音调控制曲线

### 3.3 训练策略

在这项工作中，我们通过采用具有两个编码器-解码器的编解码语言模型的分层方法来实现条件编解码语言模型的功能。这两个模型采用不同的建模方法：自回归和非自回归。我们首先应用自回归编解码语言模型从语音样本的第一个量化器生成离散代码。接下来，我们利用非自回归编解码语言模型来预测剩余的离散代码。我们可以将我们的模型表示为$\theta = \theta_{ar}, \theta_{nar}$，其中$\theta_{ar}$和$\theta_{nar}$分别表示自回归编解码语言模型和非自回归编解码语言模型的参数。

(1)自回归编解码器

在我们的方法中，我们采用自回归编解码语言建模来从语音样本的第一个量化器生成离散代码。公式为：

$$p(c^y_{:,1}|c^x_{:,1}, p, z^{(p)}; \theta_{ar}) = \prod_{t=0}^{T} p(c^y_{t,1}|c^y_{1:t-1}, c^x_{:,1}, p, z^{(p)}; \theta_{ar}) \quad (1)$$

其中$c_{:,1}$表示来自第一个量化器的离散代码，x表示语音样本，p为x对应的音素转换。通过利用自回归建模，我们可以有效地处理韵律建模过程中语音长度的变化。我们通过使用基于Transformer的编码器-解码器架构来实现自回归编解码语言模型。编码器输入是$c^x_{:,1}$和p和$z^{(p)}$的串联。

(2)非自回归编解码器

我们采用非自回归编解码语言建模来从语音样本的第2-8个量化器生成剩余的离散代码。公式为：

$$p(c^y_{:,j}|c^y_{:,j-1}, c^x_{:,j}, p, z^{(p)}; \theta_{nar}), j \in [2, 8] \quad (2)$$

其中$c_{:,j}$表示来自j的离散代码，j为第j个量化器，$\theta_{nar}$表示非自回归编解码语言模型的参数。我们同样采用基于Transformer的编码器-解码器模型作为我们的非自回归编解码语言模型，除了解码器的自注意力掩码之外，它与我们的自回归编解码器语言模型的架构类似。我们采用$c^x_{:,j}$和p和$z^{(p)}$的串联作为编码器输入。

自回归模型和非自回归模型的结合能够使合成的语音在质量和推理速度之间取得一个良好的平衡。一方面如果只用非自回归的话，模型不知道合成语音的长度应该多长，而由于自回归模型的输入序列最后有特殊的$< eos >$标记，因此自回归模型对声学序列长度的预测具有灵活性，使用自回归模型能得到语音的长度。由于第一层量化器已经得到语音长度了，那后面的量化器就可以只用非自回归，因为非自回归模型的推理速度优于自回归模型，这样就把时间复杂度从O(T)降到了O(1)。

### 3.4 音频合成

给定经过速度控制的语音样本$x = (x_1, x_2, ..., x_n)$及其对应的音素转换$p = (p_1, p_2, ..., p_n)$和音调嵌入序列$z^{(p)} = (z_1^{(p)}, ..., z_T^{(p)})$。首先，神经音频编解码器的声学量化器将每个语音样本编码


(c) 2024 中国中文信息学会计算语言学专业委员会

为离散声学代码$c^x = E_{codec}(x)$，其中$c^x \in \mathcal{N}^{n^2 \times 8}$,$n^2$表示下采样的语音长度。然后，音频合成的过程可以看作是使用多项式采样策略对

$$c^{\hat{y}} \sim p(c^y | c^x, p, z^{(p)}; \theta) \tag{3}$$

进行采样，其中$\theta$表示条件编解码语言模型的参数，$c^{\hat{y}}$是生成的离散代码序列。最后，我们再使用神经编解码器的解码器合成语音波形$\hat{y}$:

$$\hat{y} = D_{codec}(c^{\hat{y}}) \tag{4}$$

## 4 实验

### 4.1 实验设置

为了实验的公平，所有语音数据均采样至24kHz。本实验分别在英语数据集和老挝语数据集上进行训练。在训练英语数据集时，自回归编解码器epoch轮数设置为20，非自回归编解码器epoch轮数设置为40。在训练老挝语数据集时，自回归编解码器epoch轮数设置为15，训练非自回归编解码器epoch轮数设置为30。Batch size大小随机设定，初始学习率设置为0.05，实验设备为ubuntu22.04，CUDA版本为11.7，python版本为3.10.14，torch版本为1.13.1。

### 4.2 数据集

我们分别在公开的英语数据集LibriTTS(Zen et al., 2019)和私有的老挝语数据集上进行训练。LibriTTS中包含大约585小时、2456个说话者的阅读英语语音。LibriTTS语料库专为语音合成研究而设计，语音都有其对应的文本，且排除了具有显着背景噪声的语音。私有的老挝语数据集包含大约50小时、380个说话者的老挝语语音，老挝语的语音由母语为老挝语的人士进行录制，每一条语音同样有其对应的文本，其训练集和验证集的大小比为4：1。

### 4.3 评价指标

主观评价部分采用两种评价指标对模型综合能力进行评价。指标一：平均意见得分(MOS,Mean Opinion Score)，又称平均满意度得分，是一种用于评估音频自然度的指标。MOS评分采取 5 个级别对被测音频的质量进行评价，1分代表很差，2分差，3分一般，4分好，5分非常好，待测音频的得分是在所有试听人员的评分上求平均得到的。指标二：比较平均选项得分(CMOS,Comparison Mean Option Score)，在CMOS测试中，试听人员每次听两个音频，并使用一个分数来评估后者与前者相比的感觉。CMOS 的范围从-3（我们的模型比基准模型差很多）到3（我们的模型比基线模型好很多），间隔为1。

客观评价部分采用两种评价指标对模型综合能力进行评价。指标一：梅尔倒谱失真(MCD,Mel Cepstral Distortion)，它表示的是转换后语音的MFCC 特征与标准输出语音的MFCC 特征的差距，MCD的值越低代表合成音频与真实音频之间的偏差越小，即模型的效果越好。指标二：基频F0的均方根误差(RMSE)，RMSE的值越低表示合成音频与真实音频的基频轮廓越接近，效果越好。

### 4.4 对比实验

为了证明本文所提方法对韵律可控的同时，也能保证语音的自然度与流畅度，本文与以下不同语音合成模型进行对比。为了验证本文方法的有效性，分别基于LibriTTS数据集和老挝语数据集，设计本文方法与韵律相关的语音合成模型进行比较的不同实验。其中，GST-Tacotron(Wang et al., 2018)与VAE-Tacotron(Zhang et al., 2019)都是仅采用文本和参考音频作为输入的模型。GST-Tacotron使用了一种“全局风格标记”的方法来实现韵律建模。VAE-Tacotron引入VAE以无监督方式来学习韵律的潜在表达，以便于韵律控制。VITS(Kim et al., 2021)是一种结合变分推理、标准化流和对抗训练的高表现力语音合成模型。VITS在隐变量上进行随机建模并利用随机时长预测器，输入同样的文本，能够合成不同韵律的语音。VALL-E(Wang et al., 2023)是本文的基准模型。

对上述几种模型合成的语音进行主观评估，进行MOS和CMOS评分。对实验结果进行客观评估，为了方便使用MCD和RMSE评估，本文所提方法合成的语音内容与基准模型合成的语音内容一致，均方根误差(RMSE)使用基频F0进行计算。英语的实验结果如表1所示，老挝语的实验结果如表2所示，其中Ground Truth为数据集中真实的音频。


(c) 2024 中国中文信息学会计算语言学专业委员会

| 模型 | MCD | RMSE | MOS | CMOS |
|---|---|---|---|---|
| GST-Tacotron | 7.77 | 57.36 | 3.82(±0.10) | -1.32 |
| VAE-Tacotron | 7.91 | 54.88 | 3.89(±0.10) | -0.80 |
| VITS | 7.40 | 50.83 | 3.93(±0.06) | -0.41 |
| VALL-E | 7.25 | 47.39 | 3.96(±0.05) | -0.30 |
| Our Model | **7.13** | **45.21** | **4.05(±0.07)** | **0.00** |
| Ground Truth | - | - | 4.50(±0.03) | +0.25 |

表 1: 英语数据集上的各项评价指标得分

分析表1结果可知，本文所提方法合成语音的MOS评分达到4.05，相比基准模型VALL-E提升了0.09，并且我们提出的方法以+0.30的CMOS优于基准模型。同时可以看出本文提出的模型在MCD指标上相比基准模型VALL-E降低了0.12，在RMSE指标上相比基准模型降低了2.18。我们的方法在4项评价指标上均是最优的。由于GST-Tacotron和VAE-Tacotron在使用较短的参考音频时无法准确复制参考音频的说话人风格，而我们的方法基于神经编解码语言模型，它只需要3秒的参考音频，就能合成出与参考音频说话风格相似的语音。因此，在参考音频较短的情况下，我们的方法合成的音频相比GST-Tacotron和VAE-Tacotron与真实音频偏差更小、更接近，听觉上更加良好。VITS采用基于随机数的时长预测器，形成随机噪声，但随机噪声无法很好地模拟真实的韵律分布，导致其在听感上不如我们的方法。VALL-E能很好地学习参考音频的音色，但其缺乏对声音的韵律建模。根据主观评价指标和客观评价指标结果，表明我们提出的方法可以根据基准模型合成更自然、更真实的语音。

| 模型 | MCD | RMSE | MOS | CMOS |
|---|---|---|---|---|
| VITS | 7.78 | 58.26 | 3.52(±0.10) | -0.30 |
| VALL-E | 7.66 | **52.15** | 3.58(±0.12) | -0.14 |
| Our Model | **7.52** | 53.81 | **3.61(±0.15)** | **0.00** |
| Ground Truth | - | - | 4.45(±0.04) | +0.40 |

表 2: 老挝语数据集上的各评价指标得分

分析表2结果可知，本文所提方法合成的老挝语语音的MOS评分达到3.61，相比基准模型VALL-E提升了0.03，并且我们提出的方法以+0.14的CMOS优于基准模型。同时可以看出本文提出的模型在MCD指标上相比基准模型VALL-E降低了0.14。虽然在客观评价指标RMSE上结果不如基准模型，但差距并不大，这是可以接受的。VITS采用基于随机数的时长预测器，形成随机噪声，但随机噪声无法很好地模拟真实的韵律分布，导致其在听感上不如我们的方法。VALL-E能很好地学习参考音频的音色，但其缺乏对声音的韵律建模。总体来说，对于老挝语而言，所提出的方法仍可以根据基准模型合成更自然、更真实的语音。

考虑到本文所使用的私有老挝语数据集时长只有约50小时，数据量的有限可能无法展现模型的优势和特点，我们将老挝语数据集扩增至100小时并再次进行实验。

| 数据量 | MCD | RMSE | MOS | CMOS |
|---|---|---|---|---|
| Our Model-50h | 7.52 | 53.81 | 3.61(±0.15) | 0.00 |
| Our Model-100h | 7.46 | 51.94 | 3.65(±0.21) | +0.18 |

表 3: 老挝语不同数据集规模上的各评价指标得分

由表3结果可以发现，数据量的增大确实能提升模型的性能，证明了本文方法的在不同数据集规模的有效性。

### 4.5 可控性测试

为了证明本文所提出方法的可控性，我们向试听人员展示了同一句英文文本的受控音频的合成结果，以及相应的控制曲线图。然后，试听人员对受控的合成音频的音调或速度变化的准确程度进行评估，评分范围为1 到5，其中1 表示“完全不准确”，3 表示“准确”，5“非常准确”。


(c) 2024 中国中文信息学会计算语言学专业委员会

每一轮包含同一合成文本的速度控制、音调控制和音调+速度控制，每位试听人员均进行6轮测试。

| | | 速度 | 音调 |
|---|---|---|---|
| 仅速度控制 | 真实曲线 | 3.30±0.13 | - |
| 仅速度控制 | 虚假曲线 | 3.16±0.10 | - |
| 仅音调控制 | 真实曲线 | - | **3.14±0.21** |
| 仅音调控制 | 虚假曲线 | - | 2.95±0.15 |
| 速度+音调控制 | 真实曲线 | **3.34±0.08** | 3.06±0.15 |

表 4: 95%置信区间下可控性测试的MOS结果

在表4中，对于单因素(仅速度或仅音调)控制，所呈现的控制曲线有15%的可能性是虚假曲线，虚假曲线即真实曲线的翻转或移位版本。将向试听人员展示虚假曲线后得到的评分作为基线评分，对比真实曲线的评分可得，在同时进行速度和音调控制时，速度变化最准确，与基线评分相比提升0.18；在进行单因素音调控制时，音调变化最准确，与基线评分相比提升0.19。这是由于评测的时候存在一些较短的音频，仅有1-2秒的时长，试听人员可能更难评估音调的可控性。在测试中，速度控制采用上升模式，音频速度会随时间逐渐变快。音调控制采用stress模式，选取一段音频区间作为音调上升的部分，因此对于短音频，音调上升部分很短，尤其在速度变化的对比下，音调在听感上变化更加不明显。而单音调控制时，没有速度的变化，试听者在听感上更能察觉音调的变化。这就导致在速度音调同时控制时，音调评分略低于仅音调控制，但仍高于基线评分。这表明我们的方法能够同时控制这两个因素，即实现了韵律的可控性。

### 4.6 定性分析

给定两个老挝语文本和一段录音，我们分别在基准模型VALL-E和我们的模型上运行推理过程两次，并在图4和图5中可视化其波形。

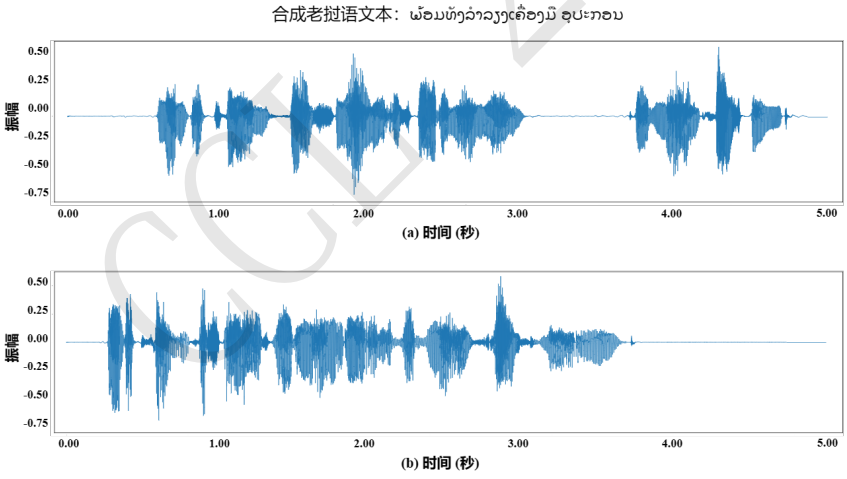



图 4: 老挝语文本1合成的声音波形

在图4中，我们观察到两个样本具有不同的持续时间。这是因为4(a)是在基准模型VALL-E上合成的样本，没有速度控制，而4(b)是经过抛物线模式的速度控制曲线修改后的样本。可以观察到，速度的差异性导致两个样本即使合成的是相同的老挝语文本，仍具有不同的持续时间。


(c) 2024 中国中文信息学会计算语言学专业委员会

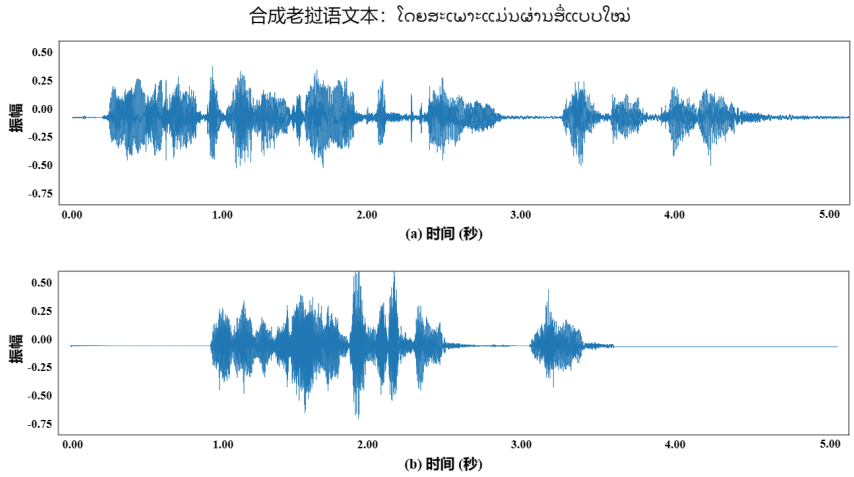



图 5: 老挝语文本2合成的声音波形

在图5中，我们观察到两个样本的振动的频率不同。类似的，5(a)是在基准模型VALL-E上合成的样本，没有音调控制，而5(b)是经过stress模式的音调控制曲线修改后的样本。可以观察到，音调的差异性导致两个样本即使合成的是相同的老挝语文本，仍具有不同的振动频率。

## 5 结论

针对现有方法中合成的语音仍然缺乏真实语音的多样性和韵律变化的丰富性的问题，本文提出基于神经编解码语言模型的老挝语韵律建模方法。利用先验速度和音调时序变化曲线建模韵律变化分布，实现对语言特征的时变控制，并且在训练中使用自回归架构和非自回归架构相结合的方式。在公开英语数据集LibriTTS和私有老挝语数据集上，主观和客观评价结果都证明，本文提出的方法能够在合成高质量音频的同时进行速度和音调控制。

## 参考文献

Anh, Nguyen Thi Ngoc, Thanh, Nguyen Tien, and others. 2022. *Development of a high quality text to speech system for Lao*, volume 1. IEEE, 2022. 2022 25th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA), pages 1–5.

Borsos, Zalán, Marinier, Raphaël, Vincent, Damien, Kharitonov, Eugene, Pietquin, Olivier, Sharifi, Matt, Roblek, Dominik, Teboul, Olivier, Grangier, David, Tagliasacchi, Marco, and others. 2023. *Audiolm: a language modeling approach to audio generation*, volume 1. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. IEEE.

Charpentier, Francis, and Stella, M. 1986. *Diphone synthesis using an overlap-add technique for speech waveforms concatenation*, volume 11. In *ICASSP'86. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 2015–2018. IEEE.

Chen, Meiying, and Duan, Zhiyao. 2022. *ControlVC: Zero-shot voice conversion with time-varying controls on pitch and speed*. arXiv preprint arXiv:2209.11866.

Défossez, Alexandre, Copet, Jade, Synnaeve, Gabriel, and Adi, Yossi. 2022. *High fidelity neural audio compression*, volume 1. arXiv preprint arXiv:2210.13438.

Dhariwal, Prafulla, Jun, Heewoo, Payne, Christine, Kim, Jong Wook, Radford, Alec, and Sutskever, Ilya. 2020. *Jukebox: A generative model for music*, volume 1. arXiv preprint arXiv:2005.00341.

Kasi, Kavita, and Zahorian, Stephen A. 2002. *Yet another algorithm for pitch tracking*, volume 1. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages I–361. IEEE.

Kim, Daegyeom, Hong, Seongho, and Choi, Yong-Hoon. 2023. *SC VALL-E: Style-Controllable Zero-Shot Text to Speech Synthesizer*, volume 1. arXiv preprint arXiv:2307.10550.


(c) 2024 中国中文信息学会计算语言学专业委员会

Kim, Jaehyeon, Kong, Jungil, and Son, Juhee. 2021. *Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech*, volume 1. In *International Conference on Machine Learning*, pages 5530–5540. PMLR.

Klimkov, Viacheslav, Ronanki, Srikanth, Rohnke, Jonas, and Drugman, Thomas. 2019. *Fine-grained robust prosody transfer for single-speaker neural text-to-speech*, volume 1. arXiv preprint arXiv:1907.02479.

Lee, Younggun, and Kim, Taesu. 2019. *Robust and fine-grained prosody control of end-to-end speech synthesis*, volume 1. IEEE, 2019. ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 5911–5915.

Mortensen, David R, Dalmia, Siddharth, and Littell, Patrick. 2018. *Epitran: Precision G2P for many languages*. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).

Skerry-Ryan, RJ, Battenberg, Eric, Xiao, Ying, Wang, Yuxuan, Stanton, Daisy, Shor, Joel, Weiss, Ron, Clark, Rob, and Saurous, Rif A. 2018. *Towards end-to-end prosody transfer for expressive speech synthesis with Tacotron*, volume 1. PMLR, 2018. International Conference on Machine Learning, pages 4693–4702.

Wang, Chengyi, Chen, Sanyuan, Wu, Yu, Zhang, Ziqiang, Zhou, Long, Liu, Shujie, Chen, Zhuo, Liu, Yanqing, Wang, Huaming, Li, Jinyu, and others. 2023. *Neural codec language models are zero-shot text to speech synthesizers*, volume 1. arXiv preprint arXiv:2301.02111.

Wang, Yuxuan, Stanton, Daisy, Zhang, Yu, Ryan, RJ-Skerry, Battenberg, Eric, Shor, Joel, Xiao, Ying, Jia, Ye, Ren, Fei, and Saurous, Rif A. 2018. *Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis*, volume 1. PMLR, 2018. International Conference on Machine Learning, pages 5180–5189.

Zen, Heiga, Dang, Viet, Clark, Rob, Zhang, Yu, Weiss, Ron J, Jia, Ye, Chen, Zhifeng, and Wu, Yonghui. 2019. *Libritts: A corpus derived from librispeech for text-to-speech*. arXiv preprint arXiv:1904.02882.

Zhang, Ya-Jie, Pan, Shifeng, He, Lei, and Ling, Zhen-Hua. 2019. *Learning latent representations for style control and transfer in end-to-end speech synthesis*, volume 1. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6945–6949. IEEE.

CCL 2024

# 基于通用依存句法的锡伯语句法树库构建研究

**周贺**
香港理工大学中文及双语学系
he.zhou@polyu.edu.hk

## 摘要

我国是一个多民族、多语种的国家，拥有丰富的民族语言资源。然而，使用人口较少、文化影响力较小的语言普遍面临语言濒危的问题，记录和保存这些语言在语言学、民族学与人类学上都具有重要意义。在本研究中，我们以我国仍在活跃使用的满通古斯语——锡伯语为目标语言，从锡伯语语法书、锡伯语报纸《察布查尔报》以及锡伯语《语文》教材中收集了 1200 个句子，以此为语料构建了一个包含词汇、形态以及依存句法信息的树库。本文详细描述了树库的构建过程，深入讨论了标注过程中遇到的难以解决的语言现象，并提出了我们的标注策略。通过标注，我们发现，随着汉语和锡伯语的深层接触，锡伯语不仅在词汇上接受了大量的汉语借词，锡伯语句子结构也受到一定程度的影响。基于所标注的锡伯语树库，我们进行了锡伯语自动句法分析实验，探讨了词、词性、字符特征以及中国少数民族语言预训练模型 CINO 对句法分析性能产生的影响。

**关键词：** 通用依存句法； 树库； 锡伯语； 满通古斯语族； 低资源语言资源

# A Dependency Treebank for Xibe based on Universal Dependencies

**He Zhou**
Department of Chinese and Bilingual Studies, Hong Kong Polytechnic University
he.zhou@polyu.edu.hk

## Abstract

China is a country of multi-ethnics and multi-languages, possessing a wealth of ethnic language resources. However, languages with fewer speakers or less influence are increasingly endangered. Therefore, the documentation of these languages is of significant importance in the fields of linguistics, ethnology and anthropology. This study focuses on the Xibe language, a Tungusic language actively used in China. We collected 1200 written Xibe sentences from a grammar book, Cabcal News and Xibe textbooks, and built a treebank covering annotations on lexical, morphological and syntactic levels. This paper outlines the detailed procedure of treebank construction, discusses challenging cases encountered in annotation, and provides justification for our current annotation decisions. Meanwhile, we also observed that due to extensive language contact with Chinese, modern written Xibe has not only enriched its lexicons from Chinese, but also shows some influence on its syntactic structures. Using the annotated treebank, we conducted dependency parsing experiments, where we investigated the effectiveness of different features including word, part-of-speech, character as well as the Chinese minority pre-trained language model (CINO).

**Keywords:** Universal Dependencies , Treebank , Xibe , the Manchu-Tungusic Languages , Low-resource Language Resources

第二十三届中国计算语言学大会论文集，第290页-第304页，太原，中国，2024年7月25日至28日。
卷1：主会论文
(c) 2024 中国中文信息学会计算语言学专业委员会

## 1 引言

锡伯语 (ISO 639-3:sjo) 是我国锡伯族仍在使用的一门语言，与满语、鄂温克语、鄂伦春语、赫哲语以及历史上的女真语构成我国境内的阿尔泰语系满-通古斯语族语言。根据 2020 年全国人口普查数据 (国务院第七次全国人口普查领导小组办公室, 2021), 锡伯族人口为 19 万 1911 人，占全国人口的 0.01%, 主要分布在东北三省和新疆维吾尔自治区 [0]。其中，使用锡伯语的人口主要集中分布在新疆伊犁哈萨克自治州察布查尔锡伯族自治县及其周边地区，懂锡伯语的人口数在 1 万至 5 万之间 (朝克, 2009; 朝克, 2014)。联合国教科文组织依据语言活力和濒危程度对世界语言进行评估，锡伯语属于严重濒危 (severely endangered) 型语言 (Moseley, 2010)。另外，锡伯语与满语关联密切，两种语言的语法结构和词汇基本一致，学术界多认为锡伯语文是满语文的继续和发展 (安俊, 1985; 赵阿平 et al., 2003)。从文字的角度来看，满通古斯语族的语言中，只有锡伯文字还有一定的使用人口，但是使用人数和范围也在缩小 (朝克, 2014)。因此，从语言学的角度记录和研究锡伯语，不仅对于保护锡伯族的语言和文化具有深远的意义，同时也有助于满语及其文献的研究。

国内外学者对锡伯语的研究主要侧重于利用田野调查收集的语料，描写和分析锡伯语的语音、词汇和语法特点，以及进行跨语言对比研究 (顾松洁, 2016)。然而，采用基于语料库或者计算语言学的方法来记录和分析锡伯语的研究则相对较少。鉴于此，本研究从语言资源建设着手，目标为锡伯语这一低资源语言构建一个包含形态与句法信息的树库。

我们选择使用通用依存句法理论 (de Marneffe et al., 2021, Universal Dependencies, 下文简称为 UD) 作为树库的标注框架。在跨语言或多语言的句法分析任务中，不同语言中的相同句法结构因采用不同的标注框架而呈现出标注差异，这严重影响了句法分析效果。为了消除这一障碍，Nivre et al. (2016) 整合了斯坦福依存句法框架 (de Marneffe and Manning, 2008), 谷歌依存句法框架 (McDonald et al., 2013), 谷歌通用词性标记集 (Petrov et al., 2012) 以及通用形态句法标记集 (Zeman, 2008)，发展出了通用依存句法 (Universal Dependencies)。UD 标注框架自提出以来已广泛应用于世界上多种不同类型的语言。这些开源的树库资源为我们锡伯语树库的标注工作提供了宝贵的参考依据。我们从锡伯语的亲属语言的树库中汲取经验，借鉴其标注方法，从而在确保标注的准确性的同时，也保持了各树库之间标注的一致性。

## 2 背景介绍

### 2.1 锡伯语的基本特点

| | | | | | |
|---|---|---|---|---|---|
| 元音 (5) | a[a] | e[ə] | i [i] | o [o] | u[u] |
| 辅音 (19) | n[n] | k [q]/[k] | g [ɢ]/[g] | h [x]/[χ] | b [p] |
| | p[p$^h$] | s [s] | š[ʂ] | t [t$^h$] | d [t] |
| | l[l] | m[m] | c[tʂ$^h$] | j[tʂ] | y[j] |
| | r[r] | f[f] | w[v] | ng[ŋ] | |
| 特定字母（10） | ck[k$^h$] | cg[g$^h$] | ch[x$^h$] | z[z] | ts[ts$^h$] |
| | dz[ts] | sy[sz] | tsy[ts$^h$z] | cy[tʂ$^h$ʐ] | jy[tʂz] |
| 满语元音 (1) | v [u] | | | | |

表 1: 锡伯语字母、其拉丁字母转写及国际音标

锡伯语与阿尔泰语系内其他亲属语言，如蒙古语族和突厥语族语言，在语言的多个层面均呈现出了共同的特征。在语音层面，锡伯语存在元音和谐现象，即一个词内的元音一般具有相同的发音特征；在形态层面，锡伯语采用黏着形态，通过在词根上附着不同的词尾来表达语法功能，而且锡伯语的黏着词缀与词根同样遵循元音和谐规律；在句法层面，锡伯语句子遵循主

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

[0] https://www.neac.gov.cn/seac/ztzl/xbz/gk.shtml

| 语料来源 | 句子数 | 总词元数 | 句长 $_{max}$ | 句长 $_{min}$ | 平均句长 | 中位数 | 方差 $\sigma$ |
|---|---|---|---|---|---|---|---|
| 语法例句 | 544 | 5 757 | 50 | 3 | 10.58 | 10 | 5.43 |
| 《察布查尔报》 | 266 | 9 644 | 90 | 3 | 36.26 | 33 | 21.52 |
| 小学课本 | 390 | 8 062 | 110 | 3 | 20.67 | 17 | 13.89 |
| 总计 | 1 200 | 23 463 | 110 | 3 | 19.55 | 13.5 | 16.65 |

表 2: 锡伯语树库语料句子的统计信息（句长的计算基于用空格分词处理之后的词元数）

语-宾语-谓语（SOV）的语序，短语遵循中心词后置（head-final）的语序。现代锡伯文文字是在圈点满文的基础上稍作修改演化而来，采用从上到下、从左到右的书写顺序。表1为锡伯语书面语字母及其拉丁字母转写和国际音标，锡伯语有 5 个元音，19 个辅音，还有 10 个用于转写外来词的特定字母以及满语元音 ᡶ（*v*）[1]。1947 年“锡索文化协会”在锡伯文字改革时，删除了满语元音 ᡶ（*v*）(安双成, 1997)，但是由于锡伯语继承了众多满语词汇，这个字母依然广泛地存在于现代锡伯语文本中。

与我国仍存的其他满通古斯语族语言 (满语、鄂温克语、鄂伦春语和赫哲语) 相比，现代锡伯语不仅同时存在书面语和口语，而且掌握两种形式的人数也多于其他语言。然而，锡伯语的书面语和口语在语音、词汇以及语法上都有较大差别 (苏承志, 1995; 佟加·庆夫, 1996; 张泰镐, 2008)。语音上，书面语词汇在口语中经历了元音音变或脱落、音节脱落以及辅音异化等变化，书面语中的多音节词在口语中逐渐双音节或单音节化；词汇上，书面语多使用规范词汇，而口语中并未使用，例如锡伯语口语中的话题标记 *da* 在书面语中并未得到使用 (Jang, 2020)；语法上，口语句子短小多变，结构简单，而书面语语法规范、准确且完整 (顾松洁, 2016)。在本研究中，我们选择锡伯语书面语作为树库的语料，主要出于以下三个原因：首先，虽然现存锡伯语口语在整体上不存在方言的区分，但新疆地区的锡伯语仍展现出细微的地区差异。国内外的研究中所使用的口语语料，例如张泰镐 (2008), Zikmundová (2013)，大多采用拉丁字母或者国际音标转写，缺少统一的文字表示，这不利于树库语料的收集和整理。其次，书面语句子严格遵守句法规则，结构规整，有助于我们准确地分析和理解锡伯语的词汇、形态及句法结构。最后，以锡伯语书面语为媒介的文学作品、语言课本和《察布查尔报》等丰富的语言资源，为语料库的构建提供了宝贵的语料。

## 2.2 锡伯语和满语的关系

锡伯族的固有语言与满语相近，锡伯人在被纳入八旗后，很快学用了满文并接受了逐步规范化了的满语。18 世纪中期，清朝为了加强西北边陲的防务，从东北抽调了一批锡伯兵西迁至新疆伊犁河谷戍边屯垦。这些锡伯人及其后代形成了今天新疆伊犁的锡伯族。在漫长的历史进程中，锡伯族与周边其他民族接触较少，居住较为集中，而且受到封建封闭式的八旗制度的约束，所以西迁的锡伯人很好地保留了满语文。至 20 世纪 40 年代，伊犁的锡伯族在旧满文的基础上进行修订和改革，形成了现代锡伯文 (佘吐肯, 2006)。从历史语言学的角度，锡伯语和满语在文字、语音、词汇和语法方面均存在着高度的一致性。现代锡伯文字是在圈点满文的基础上稍作修改演变而来的；在满语和锡伯语的基本词汇中，同源词达到 90% 以上，而且二者的语法结构也基本一致 (朝克, 2014)。国内外学者对于锡伯语的独立地位持有不同的看法，本文对此不做深入探究，主要采用佘吐肯 (2006) 的观点，认为锡伯语是一门独立的语言，它与满文具有源流的关系，在继承满文的基础上，继续丰富和发展了满文。基于此，在树库标注的过程中，除了锡伯语的参考资料，我们还可以参考满文语言材料，如《新满汉大辞典》(胡增益, 2020)，Gorelova 编写的满文语法 *Manchu Grammar* (Gorelova, 2002)，Jerry Norman 编著的满-英词典 *A Comprehensive Manchu-English Dictionary* (Norman, 2020) 以及 Roth Li 编写的满文读本 *Manchu-A Textbook for Reading Documents* (Li, 2000) 等。


(c) 2024 中国中文信息学会计算语言学专业委员会

```
# sent_id = grammarbook_sjo_p1_9
# text = ᠮᠢᠨᠢ ᠠᠮᠠ ᠤᠷᠤᠮᠴᠢ ᠴᠢ ᠶᠠᠪᠤᠬᠠ ᠃
# text[phon] = mini ama urumci ci yabuha .
# text[chn] = 我的爸爸去了乌鲁木齐。
1  ᠮᠢᠨᠢ  ᠮᠢᠨᠢ  PRON   _  _                        2  nmod:poss  _  Translit=mini
2  ᠠᠮᠠ   ᠠᠮᠠ   NOUN   _  _                        5  nsubj      _  Translit=ama
3  ᠤᠷᠤᠮᠴᠢ ᠤᠷᠤᠮᠴᠢ PROPN  _  _                        5  obl        _  Translit=urumci
4  ᠴᠢ    ᠴᠢ    ADP    _  Case=Abl                   3  case       _  Translit=ci
5  ᠶᠠᠪᠤᠬᠠ ᠶᠠᠪᠤ   VERB   _  Tense=Past|VerbForm=Fin  0  root       _  Translit=yabuha
6  ᠃     ᠃     PUNCT  _  _                        5  punct      _  Translit=.
```

图 1: 用 CoNLL-U 格式表示的依存结构例句

## 3 数据收集及预处理

### 3.1 语料收集

我们为锡伯语树库共收集了 1200 个句子，其中从《锡伯语语法通论》(佘吐肯, 2009) 收集了 544 个语法例句（共计 5757 个词元），从《察布查尔报》2019 年 1 月至 3 月的 9 期报纸中收集了 266 句新闻语料（共计 9644 个词元），另外从锡伯文小学课本《ᠨᠢᠶᠠᠮᠠᠩᠭᠠ ᠭᠢᠰᠤᠨ》(*niyamangga gisun*, “语文”) 第 3 册至第 6 册 (何文勤, 2006) 收集了 390 句（共计 8062 个词元）。我们统计了每一部分句子的句子数和句长等信息，如表2所示。在语料收集过程中，考虑到诗歌和文学语体的特殊性，我们在这一版本的树库中暂未收录。

### 3.2 数据预处理及标注

在 UD 中，每一个句子都用 CoNLL-U 格式 (Buchholz and Marsi, 2006) 来表示（见图1)。一个 CoNLL-U 格式的句子包括三类信息：(1) 以 # 标记开头的行存储句子的元信息，用来记录句子在树库里的编号（sent_id)，锡伯语文本（text)，句子的拉丁转写（text[phon]）以及句子的中文或英文翻译。(2) 句子以空格切分成词元（token)，每一个词元标注其词源编号（ID)，词形（FORM)，原形（LEMMA)，通用词性（UPOS,universal part-of-speech)，扩展词性（XPOS,extended part-of-speech)，形态特征（FEATS)，依存词（HEAD)，依存关系（DEPREL)，增强依存关系（DEPS）以及其他自定义信息（MISC)。10 个字段之间用制表符分隔开，无值字段用下划线表示。(3) 句子和句子之间的空行。

我们自动获得每一个锡伯语句子的元信息。锡伯文文本首先根据表1通过 python 脚本自动转写成拉丁形式，然后将其连同其拉丁转写及中英文翻译转换为 CoNLL-U 格式。对于每一个词元的信息，我们采用手工标注的方式。标注由两名标注者完成，其中第一名标注者完成了 464 句语法例句、新闻和小学教材的全部句子，第二名标注者完成了 80 句语法例句。由于两名标注者均为语言学专业的非锡伯语母语者，为了确保标注质量，标注和校对的过程中标注者不仅参考了锡伯文满文语法，还通过咨询锡伯语母语者获得句法知识。另外，标注工作借助 UD Annotatrix (Tyers et al., 2017) 和 conllueditor (Heinecke, 2019) 工具来提高标注效率。标注完成后，我们统计了锡伯语树库中所涉及的词性、形态和依存关系，详细数据列于表B.1、表C.1和表D.1中。

图1为一个标注完成的实例，其中对每一个词元，我们标注其原形、通用词性、形态信息，为方便阅读，我们将每一个词的拉丁转写存储在每一行的最后一列。例如，这个句子的谓语动词 ᠶᠠᠪᠤᠬᠠ（*yabuha*）的索引号为 5，是动词 ᠶᠠᠪᠤ（*yabumbi*，“行走”）的一般过去时形式，因此它的词性为 VERB, 形态特征标注为过去时（Tense=Past）限定动词（VerbForm=Fin)。作为句子的谓语，它的依存词为索引号为 0 的根节点，依存关系为 root。另外，最后一列为该词的拉丁转写（Translit=yabuha)。

## 4 树库标注中的难点

在树库标注的过程中，UD 的标注规范基本上可以覆盖锡伯语的形态和句法特征。但是，如第2节所提到的，现代锡伯语在继承了满语的基础上继续丰富和发展，随着锡伯族与其他民族的

[1]为了排版方便，正文中的锡伯语文本水平显示。


(c) 2024 中国中文信息学会计算语言学专业委员会

| 分写形式 | ᠪᠣ᠊ᠤ᠊ᠭᠤᠸᠠᠨ᠊ | ᠫᠠᠨ᠊ᠭᠤ᠊ | ᠪᠢᠶᠠᠨ᠊ᠫᠣᠣ᠊ | ᠭᠤᠩ᠊ᠬᠡ᠊ᠴᠭᠣ᠊ | ᠴᠢ᠊ᠶᠡ᠊ |
|---|---|---|---|---|---|
| 连写形式 | ᠪᠤᠤ᠊ᠭᠤᠸᠠᠨ᠊ | ᠫᠠᠨ᠊ᠭᠤ᠊ | ᠪᠢᠶᠠᠨ᠊ᠫᠣᠣ᠊ | ᠭᠤᠩ᠊ᠬᠡ᠊ᠴᠭᠣ᠊ | ᠴᠢ᠊ᠶᠡ᠊ |
| 拉丁转写 | bo u guwan | pan gu | biyan poo | gung he cgo | ci ye |
| 中文 | “博物馆” | “盘古” | “鞭炮” | “共和国” | “企业” |

表 3: 锡伯语汉语借词示例

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1-3 | ᠪᠤ᠊ᠭᠤᠸᠠᠨ᠊ | | | | | | | | |
| 1 | ᠪᠤ | ᠪᠤ | X | _ | _ | 10 | nsubj | _ | Translit=bo |
| 2 | ᠤ᠊ | ᠤ᠊ | X | _ | _ | 1 | flat | _ | Translit=u |
| 3 | ᠭᠤᠸᠠᠨ᠊ | ᠭᠤᠸᠠᠨ᠊ | X | _ | _ | 1 | flat | _ | Translit=guwan |

图 2: 多词元词的标注示例，以 ᠪᠤ᠊ᠭᠤᠸᠠᠨ᠊, *bo u guwan*, “博物馆”为例

交流日益频繁，锡伯语与汉语、维吾尔语和哈萨克语等语言的接触也逐渐增多。语言接触的一个突出表现是，锡伯语从汉语中借来了大量的词汇，同时，一些句子结构也在汉语翻译的影响下发生了细微的变化。我们选取标注中遇到的特殊语言现象，在详细分析的基础上决定标注方法。

### 4.1 由汉语借词产生的多词元词

根据 UD 标注规范，依存关系应建立在词与词之间，而这里的“词”是指句法意义上的词，并非正字法意义上的词。例如，法语的 au 是介词 à 和阳性冠词 le 的缩合形式，au 在正字法意义上算作一个词，然而从句法的角度应该算作两个词，因此在 UD 标注中，这类缩合冠词被拆分为两个词，au 被当做一个多词词元（Multi-word Token）。与此相反，锡伯语中出现了需要将几个词元合并为一个词的特殊现象，这类词以汉语借词为主，我们称之为多词元词（Multi-token Word）。

表3列出了 5 个汉语借词，锡伯语从汉语借词的方式是将汉语音节直接转写为发音相同或相似的一个或多个锡伯语音节，书写时每一个汉语音节对应的锡伯语音节写成锡伯语的一个“词”。在遵循锡伯文正字法 (新疆维吾尔自治区民族语言文学工作委员会, 1992) 的前提下，多词元词也可以连写，例如 ᠫᠠᠨ᠊ᠭᠤ（*Pan Gu*,“盘古”）也可以写作 ᠫᠠᠨᠭᠤ。然而，ᠪᠤ᠊ᠭᠤᠸᠠᠨ᠊（*bo u guwan*，“博物馆”）只有分写形式而不能连写，因为锡伯文中元音 ᠣ（*o*）与 ᠤ（*u*）没有连写形式。

在语料预处理阶段（见第3.2节），锡伯语句子按照空格切分成了词元串。词元串中既有锡伯语的固有词，也有汉语借词的不成词语素。图2为 ᠪᠤ᠊ᠭᠤᠸᠠᠨ᠊(*bo u guwan*) 的切分和标注。在标注这类多词元词时，我们在第一个词元前插入一行，标注多词元词的前后边界（1-3）和词形，其余字段为空。由于构成该词的三个词元在锡伯语中均不能单独成词，我们无法确定它们的具体词性类别，因此它们的词性确定为 X。根据 UD 对外来词或短语的标注方法，我们把多词元词的内部结构标注为一个平面结构，将第一个词元作为中心词，其后的两个词元分别依存于第一个词，依存关系为 flat。这样的标注方式取消了多词元词内部的层级结构，不会与短语结构相混淆，也不会影响句法层级结构。

### 4.2 话题标记 ᠣᠴᠢ*oci* 的特殊用法

在锡伯语书面语中，论元后一般用“格”来表示其句法功能，主格为零标记。但是，名词谓语句中的主语和名词性谓语之间有时会出现话题标记。ᠣᠴᠢ（*oci*）是锡伯语书面语中话题标记之一，它的基本功能是标记一个句子的主题。例如，在图3中，ᠣᠴᠢ（*oci*）位于主语 ᠪᠧᠵᠢᠩ（*beijing*,“北京”）之后，名词短语 ᠮᠤᠰᠧ ᠭᠤᠷᠤᠨ ᠢ ᠭᠧᠮᠤᠨ ᠬᠧᠴᠧ（*musei gurun i gemun hecen*，“我们国家的首都”）之前，其功能是用来标记话题。母语者经常将 ᠣᠴᠢ（*oci*）等同于现代汉语系动词“是”。然而在句法功能上，锡伯语 ᠣᠴᠢ（*oci*）与韩语的은/는（*eun/neun*）、日语的は（*wa*）更相近，与现代汉语中的“是”并不对应。因为“话题”是语用学概念，而且并非所有的语言都是话题优先型语言，UD 的表层标注中并未包含对句子信息结构的标注。因此，我们参考 UD 中其它话题优先型语言的标注方法，如韩语和日语，将话题标记当做特殊的主格标记，ᠣᠴᠢ（*oci*）的词性为后置词（ADP），


(c) 2024 中国中文信息学会计算语言学专业委员会

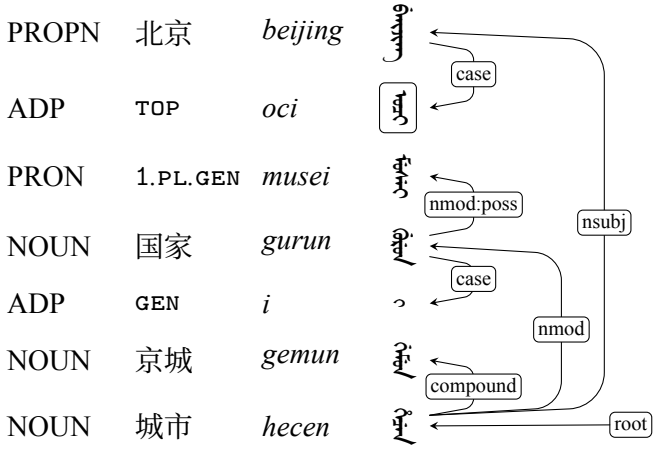



图 3: 依存结构树“北京是我们国家的首都”

它依存于主语，二者之间的依存关系为 case。

(1) a. **是**有利于群众的事他都肯干。（取自《现代汉语词典》）

b. uranda-ra jilgan be [ *pro* oci šan ]$_{clause}$ teni donji-mbi
发响-PTCP.IPFV 声音 ACC [ *pro* COP 耳朵 ]$_{clause}$ 就 听见-PRS
“发出的声响是耳朵就能听见” [2]

现代汉语中“是”除了做系动词之外，还可以用在名词前，表示“凡是”的意思，如例 (1a)。由于锡伯语母语者经常将 ᠣᠴᡳ（*oci*）直接等同于现代汉语“是”，“是”的这一功能也被锡伯语借用。例 (1b) 取自第五册《ᡧᠠᠨ ᠪᡝ ᡤᡳᡩᠠᠮᡝ ᡥᡠᠨᡤᡤᠣᠨ ᡥᡡᠯᡥᠠᠮᠪᡳ》（*šan be gidame honggon hvlhambi*，“掩耳盗铃”）。这句话中 ᠣᠴᡳ（*oci*）位于 ᡧᠠᠨ（*šan*，“耳朵”）之前，意为“凡是耳朵就可以听到发出的声响”。ᠣᠴᡳ ᡧᠠᠨ（*oci šan*）位于句子的宾语和谓语动词之间，我们将其分析为一个嵌入在主句中的主语脱落（*pro*-drop）的小句。这样，ᠣᠴᡳ（*oci*）的功能是做小句内的系动词。

图4是该句的依存结构分析，ᠣᠴᡳ（*oci*）依存于 ᡧᠠᠨ（*šan*，“耳朵”），二者的依存关系为 cop（系动关系）。小句依存于句子的谓语动词 ᡩᠣᠨᡤᡳᠮᠪᡳ（*donjimbi*，“听到”），是主句的副词性从句（advcl）。由于采用了从汉语到锡伯语逐词翻译的方式，类似例（1b）的句子忽视了锡伯语的自然表达习惯和句法结构。这不仅反映了翻译过程中的一些问题，同时也揭示了汉语对锡伯语在一定程度上的影响。

### 4.3 ᠨᡳᠩᡤᡝ *ningge* 的两种用法

ᠨᡳᠩᡤᡝ（*ningge*）是属格标记ᡳ（*i*）/ᠨᡳ（*ni*）和名词化词缀ᡤᡝ（*-ngge*）相结合而形成的。在我们收集的语料中，ᠨᡳᠩᡤᡝ（*ningge*）主要出现在两种句法环境中。

(2) a. ere bithe oci sefu i bithe .
这 书 TOP 老师 GEN 书 .
“这本书是老师的书。”

b. ere bithe oci sefu i **ningge** .
这 书 TOP 老师 GEN PTCL.POSS .
“这本书是老师的。”

c. ere bithe oci **miningge** .
这 书 TOP 1SG.GEN.POSS .

[2]语法注释释义详见表A.1


(c) 2024 中国中文信息学会计算语言学专业委员会

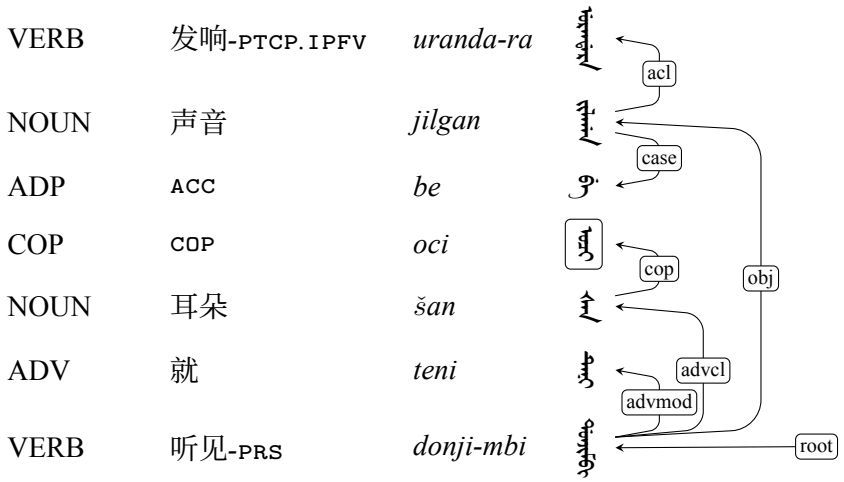



图 4: 依存结构树“发出的声响是耳朵就能听见”

“这本书是我的。”

第一种情况，ᠨᠢᠩᠭᠡ（*ningge*）附加名词或代词后，做表示领属关系的助词，替代由名词/代词和中心语所组成的名词短语 (胡增益, 2020)。例（2a）是一个名词谓语句，谓语是一个表领属关系的名词短语 ᠰᠡᠭᠦ ᠶᠢ ᠪᠢᠲᠡ（*sefu i bithe*，“老师的书”），锡伯语用属格ᠶ（*i*）标记“老师”和“书”之间的领属关系。（2b）是（2a）的省略形式，因为主语中已经提到过“书”，有时为了避免重复，中心词再次提及时可以省略，用 ᠨᠢᠩᠭᠡ（*ningge*）做领属关系助词来补足省略的成分，与前词构成表领属关系的名词短语的省略形式，语义上表示抽象概念的“东西”。例（2b）中，中心语 ᠪᠢᠲᠡ（*bithe*，“书”）省略，ᠨᠢᠩᠭᠡ（*ningge*）是属格ᠶ（*i*）和名词化词缀 ᠩᠭᠡ（*-ngge*）结合形式，与 ᠰᠡᠭᠦ（*sefu*，“老师”）一起构成表领属关系的名词短语，ᠰᠡᠭᠦ ᠨᠢᠩᠭᠡ（*sefu ningge*）意为“老师的（东西）”。当领属性名词短语的修饰语是人称代词，人称代词采用属格形式，例如例 (2c)，ᠨᠢᠩᠭᠡ（*ningge*）附加在形容词性物主代词之后，将整个短语变为名词性物主代词 (佘吐肯, 2009)。

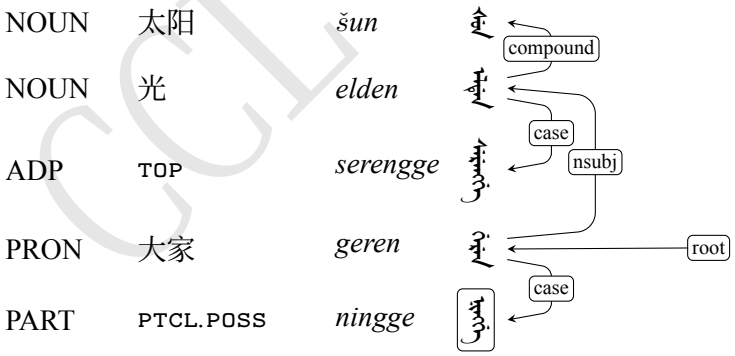



图 5: 依存结构树“阳光是大家的”

图5例句取自锡伯《语文》第六册《ᠰᡠᠨ ᡝᠯᡩᡝᠨ》(*šun elden*，“阳光”)，句中名词性谓语 ᠭᡝᠷᡝᠨ ᠨᠢᠩᠭᠡ（*geren ningge*，“大家的”）可以理解为“大家的阳光”或“大家的东西”，ᠨᠢᠩᠭᠡ（*ningge*）的作用是补足省略的中心语。在标注中，ᠨᠢᠩᠭᠡ（*ningge*）的词性为 PART（助词）, 它依存于前面紧邻的名词或代词，依存关系为 case，表明该助词赋予前面名词领属格。这种标注方式将 ᠭᡝᠷᡝᠨ（*geren*，“大家”）标注为名词短语的中心词，从语义上看不尽合理，主要是为了遵守 UD 的标注规则。UD 标注体系为了更好地捕捉不同语言之间平行性和相似句法结构，规定中心词一般为实词，虚词只能依存于实词。

另外，锡伯语书面语中，ᠨᠢᠩᠭᠡ（*ningge*）还常位于做主语的名词短语或形动词短语之后，做体词化标记（substantivizer） (Gorelova, 2002)，提示句子的话题。如图6所示的一个取自《察布


(c) 2024 中国中文信息学会计算语言学专业委员会

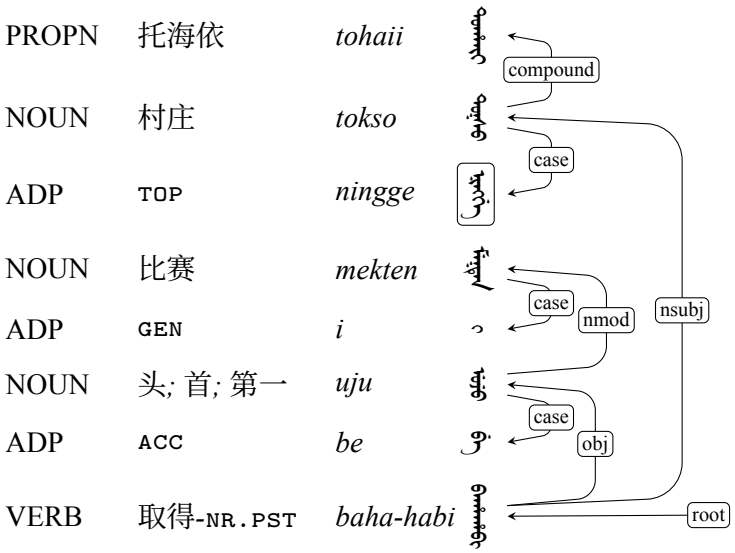



图 6: 依存结构树“托海依村获得了比赛的第一名”

查尔报》的例句，句中 ᠨᠢᠩᠭᠡ（*ningge*）在主语 ᠲᠣᠬᠠᠶ ᠲᠣᠭᠰᠣ（*tohaii tokso*，“托海依村”）之后，从句法和语义的角度来分析，我们将其分析为话题标记，它的词性为后置词（ADP），依存于前面的主语，二者依存关系为 case，表明该词赋予前面名词短语特殊的主格（同第4.2节 ᠤᠴᠢ（*oci*）的标注方法）。

### 4.4 无中心语的关系子句

关系子句（Relative Clause）的句法功能是做名词的修饰语。不同的语言中，关系子句和中心词的相对位置有前后之分，例如，英语的关系子句由关系代词引导，位于中心词之后；汉语的关系子句位于中心词之前，二者之间用“的”字连接。根据中心词在关系子句中可充当的成分，关系子句可以分为有空缺的关系子句和无空缺的关系子句，有空缺的关系子句又包括主语位置空缺和宾语位置空缺两类。例如，(3a) 中，中心语位置上的“人”是关系子句谓语“看见”的主语，而关系子句的主语位置为空，因此被称为主语空缺的关系子句；类似地，(3b) 为宾语位置空缺的关系子句。无空缺的关系子句是指中心词与关系子句中的主语和宾语均无对应关系，如 (3c)，关系子句不缺少主语或宾语，中心词“时间”也不与之相对应。在 UD 的标注体系中，关系子句的谓语动词依存于关系子句修饰的中心词。在无空缺的关系子句中，从句的谓语与中心词之间的依存关系使用 acl（修饰名词的从句）来标记；而在有空缺的关系子句中，则用 acl:relcl 来特殊标记。

(3) a. 主语空缺的关系子句：[昨天 $t_i$ 看见他的]$_{RC}$ 人 $_i$
b. 宾语空缺的关系子句：[昨天他看见 $t_i$ 的]$_{RC}$ 人 $_i$
c. 无空缺的关系子句：[我写文章的]$_{RC}$ 时间

锡伯语关系子句置于中心词之前，从句中的谓语动词采用形动词形式，而且从句的主语由属格标记，以与主句的主语进行区分。锡伯语形动词由动词词根附加形动词词尾构成，形动词词尾有完成体（4a）、非完成体（4b）以及进行体（4c）之分，其中完成体和非完成体后缀的选择基于动词词根由元音和谐律决定。

(4) a. 完成体后缀：ᠬᠠ（*-ha*），ᠬᠡ（*-he*），ᠬᠣ（-ho）
b. 非完成体后缀：ᠷᠠ（*-ra*），ᠷᠡ（*-re*），ᠷᠣ（*-ro*）或 ᠺᠠ（*-ka*），ᠺᠡ（*-ke*），ᠺᠣ（*-ko*）
c. 进行体后缀：ᠮᠠᠬᠠ（*-maha*）

图7例句中，“我爸爸种的玉米”是被关系子句修饰的名词短语，从句的主语是 ᠮᠢᠨᠢ ᠠᠮᠠ（*mini ama*，“我的爸爸”），后有属格标记；从句的谓语 ᠲᠠᠷᠢᠬᠠ（*tariha*，“种植”）是在词根 ᠲᠠᠷᠢ（*tari-*）上


(c) 2024 中国中文信息学会计算语言学专业委员会

附着完成体形动词词缀 ᠬᠠ（*-ha*）构成的形动词，修饰中心语 ᠪᠣᠯᠢᠮᠣ（*bolimo*，“玉米”）。从句内，ᠲᠠᠷᠢᠬᠠ（*tariha*，“种植”）前的宾语位置为空，中心语 ᠪᠣᠯᠢᠮᠣ（*bolimo*，“玉米”）是其语义上的宾语，因此例7符合宾语空缺的关系子句的界定。因此，根据标注规范，ᠲᠠᠷᠢᠬᠠ（*tariha*，“种植”）依存于 ᠪᠣᠯᠢᠮᠣ（*bolimo*，“玉米”），二者之间有 acl:relcl 的依存关系。

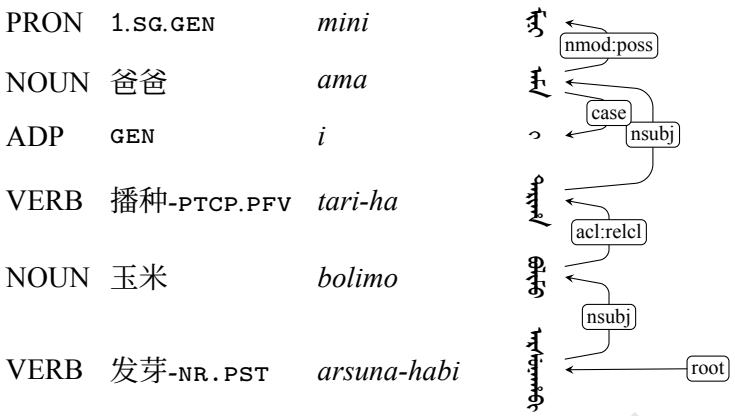



图 7: 依存结构树“我爸爸种的玉米发芽了”

除此之外，锡伯语还存在一种较为特殊的无中心词的关系子句。关系子句所修饰的中心词被，但是关系子句的谓语动词词形需要在形动词基础上再加上名词化词尾 ᠩᠭᠡ（*-ngge*），变成名动词。该词尾将形动词体词化，可以表达：（1）动作的抽象概念；（2）具有物质或非物质性质的客体；（3）行为的主体 (Gorelova, 2002)。图8所示例句中，ᠪᠡᠶᠡ ᠸᠠᠯᠢᠶᠠᠲᠠᠢ ᠬᠤᠯᠬᠠ ᠪᠡ ᠵᠠᠪᠠᠬᠠᠩᠭᠡ（*beye waliyatai hvlha be jafahangge*，“舍身擒拿盗贼的”）是一个省略了中心词 ᠨᠢᠶᠠᠯᠮᠠ（*niyalma*，“人”）的关系子句，此处名词化词尾 ᠩᠭᠡ（*-ngge*）表示“抓”的行为主体。所以，在句法层面，这种关系子句没有显性的中心词，但是在语义层面中心词隐性地存在。在标注中，这类关系子句没有用到表示关系子句的依存关系标记，名词化了的关系子句充当句子的主语。

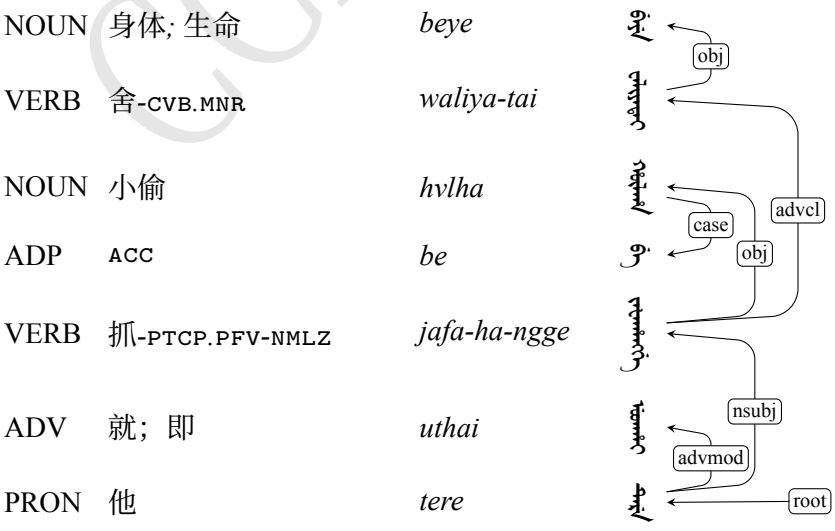



图 8: 依存结构树“舍身擒拿盗贼的就是他”


(c) 2024 中国中文信息学会计算语言学专业委员会

## 5 句法分析实验

### 5.1 数据切分及实验设置

由于树库的规模较小，为了可以更加全面地评估句法分析的性能，我们采用三折交叉验证的方法。如第3节所述，锡伯语树库的句子包括语法书例句、新闻与语文课本句子，且 3 种类型的语料数量并不均衡。为了保证每一折语料分布的均衡性，我们使用分层抽样的方法将句子按照语料体裁平均分成三折，每折 400 句。在实验中，每一折都用作训练、验证和测试一次，从而使每一个句子都可以参与一次训练和测试。

我们使用基于双仿射依存句法分析 (biaffine dependency parsing) 模型的句法分析器 MultiParser[3] (Sayyed and Dakota, 2021)。为了探究词性和字符信息是否可以有效地提高句法分析的准确率，除了词向量，我们还加入词性和字符向量来增强词的表示。此外，我们还使用从少数民族多语言模型 CINO (Yang et al., 2022, CINO-large-v2) 提取的词向量。

在测试阶段，我们计算 UAS(Unlabeled Attachment Score) 和 LAS(Labeled Attachment Score)。UAS 计算参与测试的句子中找到正确的支配词 (head) 的词所占的比例，LAS 计算测试集中既找到正确的支配词又正确地标记了二者之间依存关系 (dependency relation) 的词所占的比例。在本文中，我们计算三折交叉实验中得到的 UAS 和 LAS 的均值。

### 5.2 实验结果及分析

| 特征 | UAS | LAS |
|---|---|---|
| `word` | 68.19 | 56.57 |
| `word+pos` | 67.19 | 55.52 |
| `word+char` | 71.65 | 62.34 |
| `word+pos+char` | 71.69 | 62.17 |
| `word+char+CINO` | **72.11** | **62.47** |

表 4: 使用不同特征输入表示的句法分析结果。其中，`word` 指词向量, `pos` 指词性向量, `char` 指字符向量。

表4为使用不同的特征组合的条件下句法分析的准确率。当仅用词向量作为输入时，句法分析获得 68.19% 的 UAS 和 56.57% 的 LAS，但是当加入词性向量时，句法分析性能下降，LAS 下降了 1.05%。究其原因，由于训练数据较少，测试数据中平均高达 46.38% 的词在训练数据中并未出现，因而模型并未充分地学习到词汇和句法信息。当我们继续加入词性信息时，词性信息并未帮助消歧，反而增加了噪音，使句法分析表现下降。当输入表示用词向量和字符向量时，LAS 提高了 5.77%(62.34 vs. 56.57)，这表明字符向量提供给句法分析有用的词汇信息以及字符之间的位置信息，有效地缓解了未登录词的问题。然而，当输入表示将词、词性和字符向量相结合时，句法分析性能并没有进一步提高，LAS 较使用词和字符向量的模型下降了 0.17%，虽然这一差异并不显著，这仍然可以表明词性信息在本实验设置条件下并未起到正面的作用。

当使用词、字符和 CINO 向量作为输入时，句法分析性能达到最佳，UAS 达到 72.11%，LAS 达到 62.47%。但是，该模型只高于使用词和字符特征的模型 0.13%(62.47 vs. 62.34)，这表明 CINO 语言模型能提供给锡伯语句法分析的信息也较为有限。CINO 模型基于多语言训练模型 XLM-R (Conneau et al., 2020)，在藏语、蒙语、维吾尔语、哈萨克语、朝鲜语、壮语和粤语上进行二次训练，其训练数据中并没有锡伯语数据，与锡伯语文字相似的是回鹘体蒙文。我们进一步对比了 Unicode 编码表[4]中蒙古文和锡伯文字母的差异，发现锡伯语只有 12 个字母与蒙古文具有相同的 unicode 编码，其余 21 个字母为锡伯语独有，是 CINO 预训练语言中所不覆盖的。因此，锡伯语句法分析性能得到轻微提升，主要归功于 CINO 模型中与蒙古语同码的 12 个字母所带来的知识。

[3] https://github.com/zeeshansayyed/multiparser
[4] https://www.unicode.org/charts/PDF/U1800.pdf


(c) 2024 中国中文信息学会计算语言学专业委员会

## 6 小结及下一步工作

在本研究中，我们基于通用依存句法，为我国资源稀缺的少数民族语言——锡伯语，手工标注了一个含有词汇、形态和句法信息的依存树库。树库目前包括从《锡伯语语法通论》、《察布查尔报》和锡伯《语文》收集的 1200 个标注完整的句子。我们详细介绍了树库的标注过程，深入分析了标注过程中遇到的较难处理的词汇和句法层面的语言现象，并提出了我们的标注策略。这些语言现象同时也反映出，与汉语的深层次接触导致锡伯语不仅在词汇层面上产生变化，锡伯语句法也产生了微妙的变化。基于所标注的锡伯语树库，我们进行了锡伯语句法分析实验，实验中我们重点研究了不同的特征对句法分析器性能的影响，发现在该实验条件下，输入字符可以为句法分析提供有用的信息，而词性会带来干扰。另外，受预训练数据和文字类型所限，锡伯语句法分析器从少数民族预训练语言模型 CINO 中受益有限。

我们的下一步工作主要包括两个方面：首先，我们将继续从《察布查尔报》和锡伯语文学作品中收集语料，扩大树库的规模；其次，手工标注句法树库是一项耗时耗力的工作，我们将以手工标注的树库为基础，继续深入研究锡伯语自动词法分析和句法分析，实现树库的自动标注。与此同时，我们也将扩大工作范围，为其他的少数民族语言构建语言资源及应用。

## 参考文献


Sabine Buchholz and Erwin Marsi. 2006. Conll-x shared task on multilingual dependency parsing. In *Proceedings of the tenth conference on computational natural language learning (CoNLL-X)*, pages 149–164.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Oline.

Marie-Catherine de Marneffe and Christopher Manning. 2008. The stanford typed dependencies representation. In *Coling 2008: proceedings of the workshop on cross-framework and cross-domain parser evaluation*, pages 1–8.

Marie-Catherine de Marneffe, Christopher Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational Linguistics*, 47(2):255–308, 07.

Liliya M Gorelova. 2002. *Manchu Grammar*. Brill.

Johannes Heinecke. 2019. ConlluEditor: a fully graphical editor for Universal Dependencies treebank files. In *Universal Dependencies Workshop 2019*, Paris.

Taeho Jang. 2020. Xibe and the manchuric languages. In Martine Robbeets and Alexander Savelyev, editors, *The Oxford Guide to the Transeurasian Languages*, pages 269–287. Oxford University Press.

Gertraude Roth Li. 2000. *Manchu: a Textbook for Reading Documents*. University of Hawaii Press.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, et al. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97.

Christopher Moseley. 2010. *Atlas of the World's Languages in Danger*. UNESCO.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: a multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portoroz, Slovenia.

Jerry Norman. 2020. *A comprehensive Manchu-English dictionary*. Brill.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2089–2096, Istanbul, Turkey, May. European Language Resources Association (ELRA).



(c) 2024 中国中文信息学会计算语言学专业委员会

Zeeshan Ali Sayyed and Daniel Dakota. 2021. Annotations matter: Leveraging multi-task learning to parse UD and SUD. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3467–3481, Online, August.

Francis Tyers, Mariya Sheyanova, and Jonathan Washington. 2017. UD Annotatrix: an annotation tool for universal dependencies. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pages 10–17.

Ziqing Yang, Zihang Xu, Yiming Cui, Baoxin Wang, Min Lin, Dayong Wu, and Zhigang Chen. 2022. CINO: A Chinese minority pre-trained language model. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3937–3949, Gyeongju, Republic of Korea, October.

Daniel Zeman. 2008. Reusable tagset conversion using tagset drivers. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*.

Veronika Zikmundová. 2013. *Spoken Sibe: Morphology of the Inflected Parts of Speech*. Karolinum Press.

何文勤. 2006. *nimangga gisun* 语文. 新疆教育出版社.

佘吐肯. 2006. 论锡伯文和满文的源流关系. 伊犁师范学院学报: 社会科学版, (4):34–39.

佘吐肯. 2009. *sibe gisun i gisun kooli hafupi leolen* 锡伯语语法通论. 新疆人民出版社.

佟加·庆夫. 1996. 现代锡伯语书面语与口语比较. 满语研究, 22(1):48–60.

国务院第七次全国人口普查领导小组办公室. 2021. 中国人口普查年鉴 *2020*. 中国统计出版社.

安俊. 1985. 锡伯语言文字乃满语满文的继续. 满语研究, 1(1):41–47.

安双成. 1997. 锡伯族与满语文. 满语研究, 25(2):18–26.

张泰镐. 2008. 锡伯语语法研究. 云南民族出版社.

新疆维吾尔自治区民族语言文学工作委员会. 1992. *nei fon sibe šu tacin gisun i arara kooli* 现代锡伯文学语言正字法. 新疆人民出版社.

朝克. 2009. 中国的濒危民族语言文字的保护及学术价值. 中国社会科学院研究生院学报, (4):120–129.

朝克. 2014. 满通古斯语族语言研究史论. 中国社会科学出版社.

胡增益. 2020. 新满汉大词典. 商务印书馆.

苏承志. 1995. 简论现代锡伯语口语和书面语的相异性. 语言与翻译, 43(3):48–56.

赵阿平, 郭孟秀, 唐戈, 吴雪娟, and 杨惠滨. 2003. 满-通古斯语族语言文化抢救调查——新疆察布查尔县锡伯族语言文化调查报告. 满语研究, 37(2):75–81.

顾松洁. 2016. 锡伯语研究综述. 满语研究, 2(2):83–87.

## 附录

### A 论文中涉及的语法注释缩写

| 缩写 | 意义 | 缩写 | 意义 |
|---|---|---|---|
| 1 | 第一人称 | IPFV | 未完成体 |
| 2 | 第二人称 | PFV | 完成体 |
| 3 | 第三人称 | PROG | 进行体 |
| ACC | 宾格 | NMLZ | 名词化标记 |
| GEN | 属格 | NR.PST | 最近过去时 |
| PTCP | 形动词 | PRS | 现在时 |
| COP | 系动词 | PST | 过去时 |
| CVB | 副动词 | POSS | 领属 |
| SG | 单数 | PTCL | 助词 |
| PL | 复数 | TOP | 话题标记 |

表 A.1: 语法注释缩写


(c) 2024 中国中文信息学会计算语言学专业委员会

## B 树库标注的词性标记及其分布

| 词性标记 | 意义 | 语法例句 | 新闻 | 小学课文 |
|---|---|---|---|---|
| NOUN | 名词 | 1344 | 3262 | 2363 |
| VERB | 动词 | 1222 | 1868 | 1575 |
| PUNCT | 标点符号 | 908 | 987 | 1185 |
| ADP | 后置词 | 615 | 1245 | 837 |
| PRON | 代词 | 439 | 190 | 297 |
| ADV | 副词 | 350 | 254 | 381 |
| ADJ | 形容词 | 326 | 662 | 556 |
| NUM | 数词 | 145 | 344 | 304 |
| PROPN | 专有名词 | 91 | 431 | 153 |
| SCONJ | 从属连词 | 80 | 7 | 35 |
| DET | 限定词 | 73 | 33 | 58 |
| PART | 助词 | 62 | 71 | 65 |
| AUX | 助动词 | 55 | 66 | 52 |
| X | 无类别 | 33 | 132 | 140 |
| CCONJ | 并列连词 | 10 | 91 | 60 |
| INTJ | 疑问词 | 4 | 0 | 1 |
| SYM | 特殊符号 | 0 | 1 | 0 |
| 总计 | - | 5757 | 9644 | 8062 |

表 B.1: 锡伯语树库的词性标记及其分布

## C 树库标注的形态特征及其分布

| 特征 | 值 | 语法例句 | 新闻 | 小学课文 |
|---|---|---|---|---|
| Abbr | Yes | 0 | 20 | 0 |
| Aspect | Imp | 407 | 1318 | 742 |
| | Perf | 143 | 240 | 371 |
| | Prog | 12 | 9 | 20 |
| Case | Abl | 45 | 73 | 48 |
| | Acc | 201 | 556 | 348 |
| | Dat | 226 | 190 | 134 |
| | Gen | 207 | 404 | 330 |
| | Ins | 0 | 12 | 14 |
| | Lat | 13 | 6 | 23 |
| | Loc | 1 | 114 | 139 |
| Clusivity | Ex | 9 | 6 | 2 |
| | In | 1 | 13 | 0 |
| Foreign | Yes | 0 | 1 | 0 |
| Mood | Cnd | 67 | 7 | 42 |
| | Des | 0 | 0 | 7 |
| | Imp | 103 | 8 | 15 |
| | Opt | 5 | 0 | 15 |
| | Sub | 4 | 1 | 1 |
| NumType | Card | 103 | 276 | 269 |
| | Frac | 0 | 2 | 0 |
| | Mult | 8 | 0 | 2 |
| | Ord | 11 | 44 | 20 |
| Number | Plur | 95 | 62 | 67 |

表 C.1: 锡伯语树库形态标注及其分布

CCL 2024

| 特征 | 值 | 语法例句 | 新闻 | 小学课文 |
|---|---|---|---|---|
| | Sing | 267 | 31 | 153 |
| Person | 1 | 141 | 38 | 76 |
| | 2 | 98 | 4 | 32 |
| | 3 | 98 | 26 | 90 |
| Polarity | Neg | 177 | 36 | 93 |
| Polite | Elev | 1 | 0 | 2 |
| Poss | Yes | 89 | 13 | 68 |
| PronType | Dem | 88 | 36 | 71 |
| | Int | 38 | 1 | 29 |
| | Prs | 340 | 69 | 199 |
| | Tot | 11 | 105 | 17 |
| Reflex | Yes | 14 | 10 | 11 |
| Tense | Past | 189 | 168 | 247 |
| | Pres | 368 | 110 | 176 |
| Typo | Yes | 0 | 5 | 0 |
| VerbForm | Conv | 414 | 852 | 715 |
| | Fin | 518 | 284 | 444 |
| | Part | 177 | 594 | 319 |
| | Vnoun | 38 | 127 | 51 |
| Voice | Cau | 40 | 98 | 73 |
| | Pass | 5 | 13 | 13 |

表 C.1: 锡伯语树库形态标注及其分布

## D 树库标注的依存关系及其分布

| 依存关系 | 意义 | 语法例句 | 新闻 | 小学课文 |
|---|---|---|---|---|
| acl | 形容词性从句 | 61 | 232 | 127 |
| acl:relcl | 关系子句 | 51 | 205 | 87 |
| advcl | 副词性从句 | 489 | 856 | 775 |
| advmod | 副词性修饰语 | 376 | 328 | 429 |
| amod | 形容词性修饰语 | 153 | 398 | 313 |
| appos | 同位语修饰语 | 13 | 78 | 39 |
| aux | 助动词 | 39 | 34 | 23 |
| case | 格标记 | 610 | 1239 | 835 |
| cc | 并列连词 | 13 | 93 | 60 |
| ccomp | 从句补足语 | 37 | 26 | 31 |
| clf | 量词 | 3 | 63 | 39 |
| compound | 复合关系 | 188 | 492 | 349 |
| conj | 并列关系 | 133 | 374 | 187 |
| cop | 助动词 | 9 | 30 | 28 |
| csubj | 从句性主语 | 11 | 0 | 19 |
| det | 限定词 | 70 | 121 | 62 |
| discourse | 话语标记 | 61 | 8 | 21 |
| fixed | 固定多词表达 | 4 | 4 | 8 |
| flat | 平面多词表达 | 31 | 152 | 53 |
| flat:name | 人名 | 2 | 176 | 63 |
| flat:num | 复杂数字 | 0 | 0 | 55 |
| mark | 标记 | 78 | 13 | 34 |
| mark:adv | 副词化标记 | 4 | 46 | 35 |
| mark:plur | 复数标记 | 0 | 5 | 1 |

表 D.1: 锡伯语树库依存关系及其分布

第二十三届中国计算语言学大会论文集，第290页-第304页，太原，中国，2024年7月25日至28日。
卷1：主会论文
(c) 2024 中国中文信息学会计算语言学专业委员会

| 依存关系 | 意义 | 语法例句 | 新闻 | 小学课文 |
|---|---|---|---|---|
| mark:rel | 形容词化标记 | 1 | 4 | 11 |
| nmod | 名词性修饰语 | 256 | 1297 | 537 |
| nmod:poss | 具有领属关系的名词性修饰语 | 89 | 26 | 72 |
| nsubj | 名词性主语 | 509 | 434 | 603 |
| nsubj:pass | 被动结构中的名词性主语 | 0 | 7 | 0 |
| nummod | 数词修饰语 | 99 | 240 | 215 |
| obj | 宾语 | 347 | 754 | 619 |
| obl | 名词性间接格 | 406 | 319 | 547 |
| obl:lmod | 方位修饰语 | 0 | 103 | 0 |
| obl:tmod | 时间修饰语 | 41 | 91 | 0 |
| parataxis | 意合关系 | 37 | 81 | 126 |
| punct | 标点符号 | 908 | 987 | 1185 |
| root | 根节点 | 544 | 266 | 390 |
| vocative | 称呼语 | 5 | 0 | 14 |
| xcomp | 控制结构、提升结构等 | 70 | 61 | 62 |

表 D.1: 锡伯语树库依存关系及其分布

CCL 2024

# 基于字节对编码的端到端藏语语音识别研究

蔡郁青[1,2]，王超[2,3]，仁增多杰[*1,2]，朱宇雷[1,2]，张瑾[1,2]，尼玛扎西[*1,2]
1.西藏大学，信息科学技术学院，拉萨，850000
2.藏语智能信息处理及应用国家重点实验室，拉萨，850000
3.青海师范大学，计算机学院，西宁，810016
yuqingcai@stu.utibet.edu.cn,wangchao980328@163.com,rzdj@utibet.edu.cn,
zhuyulei0808@163.com,jinzhang@stu.utibet.edu.cn,niqiongda@163.com

## 摘要

针对藏语端到端语音识别研究中存在的建模单元不统一和识别效果不理想的问题，本文提出了一种BPE-Conformer-CTC/Attention端到端藏语语音识别方法。首先，该方法采用了字节对编码算法进行语音建模，通过反复合并出现频率最高的字符对，将文本分割成易于管理、有意义的单元，平衡建模单元的粒度，从而解决藏语语音识别中建模单元不统一的问题。其次，使用了Conformer编码器，有效地融合了音频序列的全局和局部依赖关系，从而增强了模型的表征能力。最后，通过CTC/Attention联合解码策略，加速了对齐和解码过程，进而提高了识别效果的准确性和效率。在开源数据集XBMU-AMDO31和TIBMD@MUC上的实验结果表明，所提出的BPE-Conformer-CTC/Attention模型分别取得了9.0%和4.6%的词错误率，相较于基线模型Transformer-CTC/Attention，词错误率分别相对降低了14.2%和30.3%。该研究方法为藏语端到端语音识别任务提供了一种有效的解决方案。

**关键词：** 藏语语音识别；端到端；字节对编码；安多方言

# End-to-End Tibetan Speech Recognition Study Based on Byte Pair Coding

**Yuqing Cai[1,2],Chao Wang[2,3],Renzeng Duojie[*1,2],Yulei Zhu[1,2],Jin Zhang[1,2],Nyima Tashi[*1,2]**
1.Tibet University,School of Information Science and Technology,Lhasa,850000
2.The State Key Laboratory of Tibetan Intelligent Information Processing and Application,Lhasa,850000
3.Qinghai Normal University,The College of Computer,Xining,810016
yuqingcai@stu.utibet.edu.cn,wangchao980328@163.com,rzdj@utibet.edu.cn,
zhuyulei0808@163.com,jinzhang@stu.utibet.edu.cn,niqiongda@163.com

## Abstract

Aiming at the problems of inconsistent modeling units and unsatisfactory recognition results in Tibetan end-to-end speech recognition research, this paper proposes a BPE-Conformer-CTC/Attention end-to-end Tibetan speech recognition method. Firstly, the method adopts the byte-pair encoding algorithm for speech modeling, which solves the problem of inconsistent modeling units in Tibetan speech recognition by repeatedly merging the most frequent character pairs, segmenting the text into manageable and meaningful units, and balancing the granularity of modeling units. Secondly, the Conformer encoder is utilized to effectively fuse the global and local dependencies of audio

*尼玛扎西（通信作者）：niqiongda@163.com，仁增多杰（通信作者）：rzdj@utibet.edu.cn
**基金项目：**新一代人工智能国家科技重大专项(2022ZD0116100)，西藏大学研究生“高水平人才培养计划”项目(2022-GSP-S096)
©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

sequences, thus enhancing the model's representational capability. Finally, the alignment and decoding process is accelerated by a joint CTC/Attention decoding strategy, which in turn improves the accuracy and efficiency of the recognition results. Experimental results on the open-source datasets XBMU-AMDO31 and TIBMD@MUC show that the proposed BPE-Conformer-CTC/Attention model achieves a word error rate of 9.0% and 4.6%, respectively, and compared to the baseline Transformer-CTC/Attention model, the word error rate is reduced by 14.2% and 30.3%, respectively. This proposed approach provides an effective solution for the Tibetan end-to-end speech recognition task.

**Keywords:** Tibetan speech recognition , End to end , Byte pair encoding , Amdo dialect

## 1 引言

近年来，端到端的自动语音识别(ASR)方法越来越受到关注 (Amodei et al., 2015; Chorowski et al., 2015; Huang et al., 2020)。与传统语音识别方法相比，端到端ASR系统将发音词典、声学模型和语言模型等模块集成到单个框架中进行联合优化，大幅降低了训练和推理的复杂度。

藏语端到端语音识别起步较晚。(黄晓辉 and 李京, 2018)提出结合RNN和CTC的方法建立藏语声学模型并进行端到端训练，以提高准确性和性能。(王松, 2019)采用LSTM-CTC端到端方法进行藏语拉萨话识别。随着Transformer(Vaswani et al., 2017)在机器翻译任务中的广泛应用，(Dong et al., 2018)首次将其引入语音识别领域，提出了Speech-Transformer模型。(Yang et al., 2020)提出了基于Transformer的藏语语音识别方法，使藏语识别错误率降至30%以下。(高耀荣 and 边巴旺堆, 2023)构建了Transformer-CTC/Attention模型，为后续藏语语音识别研究奠定了基础。

尽管端到端藏语语音识别取得了一定进展，但仍存在一些问题。相较于主流语言，藏语端到端识别效果较差。处理具有复杂构字特点的藏语时，建模单元存在不统一和重复实验的问题。同时，未知词汇的处理问题也一直制约着藏语识别性能的提升。

针对上述问题，本文以安多方言为研究对象，结合CTC(Graves, 2006)自动对齐和Attention上下文建模的优点，探索了基于CTC/Attention联合解码的Transformer语音识别模型。为解决Transformer捕捉局部信息能力有限的问题，提出使用Conformer作为编码器，构建了Conformer-CTC/Attention模型。在建模单元选择上，采用BPE算法进行语音建模，以解决构件、字丁和音节等单元的争议，同时通过BPE的灵活组合一定程度上能解决未知词汇问题。

## 2 建模单元的选择

对于语音识别模型来说，建模单元的选择是影响模型识别性能的重要因素之一。不同的建模单元所需的训练数据规模和模型训练难度存在差异，最终会对模型的识别精度产生影响。在藏语语音识别研究领域，由于研究者使用的训练语料规模不同，目前采用何种粒度的建模单元尚未统一。与此同时，语音识别过程中未知词汇的问题也一直没有得到有效解决。

### 2.1 藏语文字构成

藏语作为汉藏语系的一种语言，包含卫藏、康巴和安多三大方言。尽管不同方言使用同一个藏语文字系统，但研究藏字的构字规则对于覆盖所有方言而言意义重大。藏语是一种拼音文字，由30个辅音和4个元音构成。一个藏字通常对应一个音节，由字丁横向叠加而成，而字丁又由构件纵向叠加而成。因此，藏语的建模单元可分为构件、字丁和音节三种，具体示例如图1所示。

### 2.2 基于字节对编码的语音建模

在端到端语音识别系统中，建模单元粒度的选择是一个需要权衡的问题。采用较小粒度的建模单元(如构件)虽然能够覆盖所有词汇、降低标注成本，但也会导致输出序列过长、模型规


(c) 2024 中国中文信息学会计算语言学专业委员会

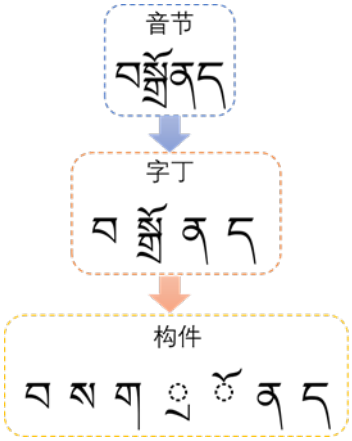



**图 1.** 藏字构成示例

模庞大、参数量剧增，从而加大了模型优化的难度，进而影响识别精度。相反，选择较大粒度的建模单元(如字丁和音节)则可以有效控制输出序列长度和模型大小，但存在词典覆盖范围有限、标注成本高昂的弊端。因此，选择适当粒度的建模单元对于平衡上述利弊至关重要，不仅能够有效控制输出序列长度、降低建模复杂度，还可以在一定程度上解决未登录词问题，提高识别性能。

针对这一问题，本文提出了使用字节对编码(Byte Pair Encoding，BPE)(Shibata et al., 1999)算法进行藏语语音建模。BPE是一种数据压缩算法，可以通过一个有限的词表来解决所有词的分词问题，同时能够较好地平衡未登录词的问题。BPE算法首先将每个藏字拆分成单个构件，然后依次用另一个字符替换频率最高的一对字符，直至达到预设的终止条件。根据训练文本，我们可以获得藏文句中各音节的BPE拆分结果，如图2所示，其中“_”表示该子词为词前缀。

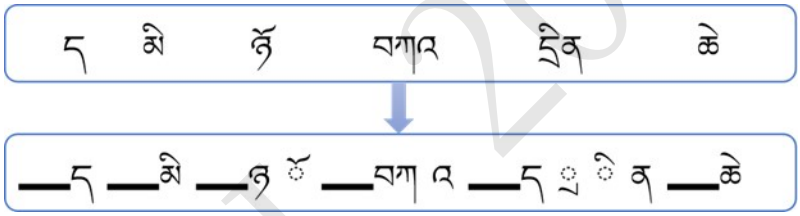


**图 2.** 藏文句BPE拆分示例

## 3 模型构建

针对藏语端到端语音识别任务，本文提出了一种BPE-Conformer-CTC/Attention混合架构。如图3所示，该架构由三个主要部分组成：Conformer编码器、CTC解码分支和Attention解码分支。首先，音频特征被Conformer编码器编码为高级语音特征序列，然后该序列并行传递给后续的CTC和Attention两个解码分支进行处理。在训练过程中，两个解码分支的损失函数按照一定权重线性组合，实现多任务联合学习。而在解码阶段，两个分支的输出概率则通过加权束搜索算法进行融合，得到最终的识别结果。

### 3.1 编码器层

本文采用Conformer(Gulati et al., 2020)作为编码器，其结构如图3所示。Conformer编码器主要由四个子模块组成，包括第一个前馈神经网络模型、多头自注意力模块、卷积模块和第二个前馈神经网络模块。不同于Transformer，Conformer中的多头注意力模块使用了相对位置嵌入编码，并采用了Swish激活函数。另一方面，卷积模块以逐点卷积(Pointwise Convolution)和门控线性单元(GLU)为起点，然后进行一维深度卷积和批次归一化(Batch norm)，这一设计有助于有效地训练深度模型。具体地，对于第$k$个Conformer模块的输入序列$x_k$，其输出$y_k$ 的计算流程如下所示：

$$x_{FFN1} = x_k + 1/2FFN(x_k) \tag{1}$$


(c) 2024 中国中文信息学会计算语言学专业委员会

$$x_{MHSA} = x_{FFN1} + MHSA(x_{FFN1}) \tag{2}$$

$$x_{Conv} = x_{MHSA} + Cona(x_{MHSA}) \tag{3}$$

$$y_k = LayerNorm(x_{Conv} + 1/2FFN(x_{Conv})) \tag{4}$$

其中$FFN$表示前馈神经网络模块，$MHSA$表示多头自注意力模块，$Conv$表示卷积模块，每个模块间都使用残差连接。

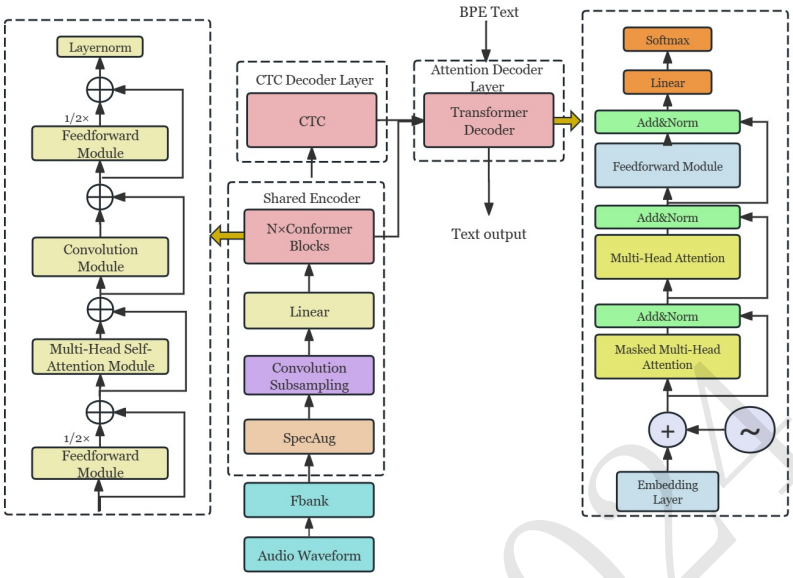



**图 3.** BPE-Conformer-CTC/Attention架构图

## 3.2 解码器层

### 3.2.1 Transformer解码器

Transformer解码器通过自注意力机制直接建模输出序列中元素之间的依赖关系，而不需要递归计算，并且Transformer架构中的多头注意力机制和位置编码有助于捕获输入和输出序列中的长依赖关系。因此，本文将Transformer解码器部分直接用作端到端语音识别模型解码器的一部分。具体应用的多头注意力机制如图4所示。

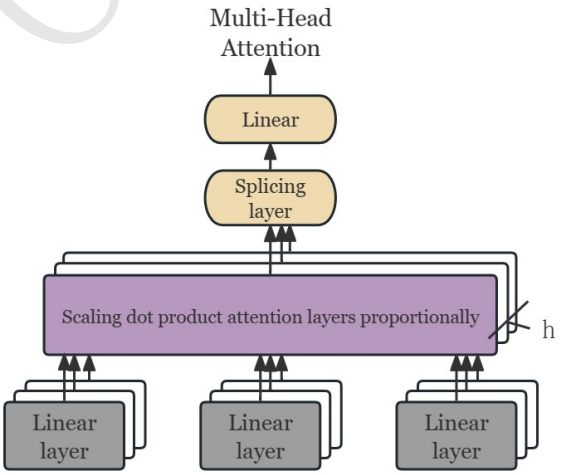



**图 4.** 缩放点积注意力流程图


(c) 2024 中国中文信息学会计算语言学专业委员会

注意力机制是基于缩放的点积注意力方程，其定义如下：

$$Attention(Q,K,V) = Softmaxa((QK^T)/\sqrt{(d_k)})V \tag{5}$$

#### 3.2.2 连接时序分类

CTC通过对目标序列和原始输入序列的对齐分布进行建模，消除了在单个声学特征建模单元和文本序列建模单元之间手动对齐的繁琐操作，从而成功实现了端到端声学模型的训练。该方法允许输出中存在重复的标签，并在这些重复标签之间插入一个特殊的空白符号，例如“_”，以形成与输入语音帧长度相匹配的CTC路径。这种灵活性使得模型能够更好地适应不同语音输入的变化和复杂性。CTC的实现基于softmax分类层，该分类层包含了所有可能类别的节点以及用于表示空白的节点。

#### 3.2.3 混合CTC/Attention模型

在处理大规模数据时，传统的Attention机制对齐方式可能会导致训练时间增加，这源于它缺乏明确的时间前后顺序依赖性。相比之下，CTC中采用的前向-后向算法则能够按照时间顺序依次高效地对齐输入序列和输出序列。为了解决Attention对齐效率较低的问题，同时提高整体训练效率，本文设计了一种混合CTC/Attention模型。在该模型中，CTC被引入作为辅助任务，用于加速对齐和解码过程．模型的整体结构示意图如图5所示。

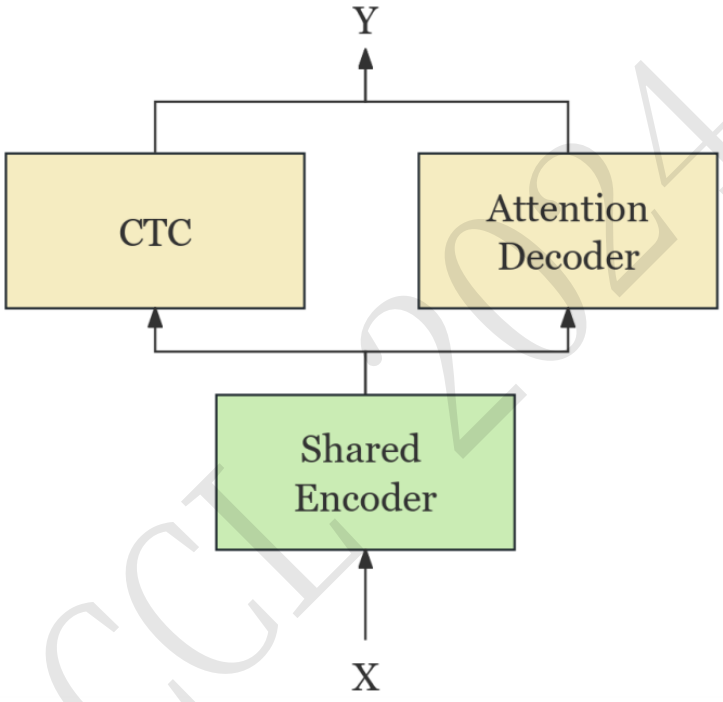



**图 5.** 混合CTC/Attention结构简图

### 3.3 损失函数

为联合优化模型，在帧级别上采用CTC损失，在标签级别上采用CE损失进行训练。本文模型损失函数的计算是将CTC损失(Lee and Watanabe, 2021)和CE损失加权相加，具体公式如下。

$$L(x,y) = \lambda CTC(x,y) + (1-\lambda)CE(x,y) \tag{6}$$

其中$x$是模型语音特征输入，$y$是相应的模型输出，以及$\lambda$是平衡CE损失和CTC损失的比重。

## 4 实验结果及分析

### 4.1 实验数据及评价指标

本文实验数据来自OpenSLR网站上开源的藏语安多数据集XBMU-AMDO31(Li et al., 2022)和藏语多方言TIBMD@MUC数据集(Zhao et al., 2020)。数据集XBMU-AMDO31由66位

母语发言人在安静的室内环境下通过手机录制，含有大约31小时的录音，22630个句子，总字数为2754。数据集TIBMD@MUC由87位母语发言人在室内环境下通过录音机、笔记本电脑和手机录制，包括卫藏方言52.53小时、康巴方言5.87小时和安多方言25.93小时，10390个句子，总字数为3006。两个数据集音频文件的采样频率均为16KHz，量化精度为16位。实验时，TIBMD@MUC数据集中仅采用25.93小时的安多方言数据，并对两个数据集的原始音频进行0.9倍和1.1倍的速度扰动将数据集扩充至3倍(Ko et al., 2015)。

采取词错误率(word error rate, WER)作为评价指标，其中S代表替换的音节数量，D代表删除的音节数量，I代表插入的音节数量，N代表总的音节数量，WER的值越小越好。具体公式如下：

$$WER = (S + D + I)/N \tag{7}$$

### 4.2 实验环境与参数

本文的深度学习框架基于Pytorch开发，编程语言为Python。具体工作站配置为：

| 环境详情 | |
|---|---|
| 操作系统 | Ubuntu18.04 |
| Linux | 4.15.0 |
| 显卡数量 | 4 |
| 显卡版本 | Tesla P100 |
| 内存 | 12G |
| Python版本 | 3.8 |
| Torch版本 | 1.13.1 |
| Cuda版本 | 11.2 |

**表 1.** 平台配置

实验中，使用80维Fbank(filter banks）作为输入特征，其中帧长为25 ms，帧移为10 ms。模型中的多头注意力头数均为4个，Encoder块有12个，Decoder块有6个，每个块的参数独立。训练中采用Swish作为激活函数，Dropout系数为0.1。使用Adam优化器，采用Warmup Step为30000的Warmup lr学习策略训练，设定学习率为0.0005，训练50轮。

### 4.3 实验分析

本文首先在XBMU-AMDO31和TIBMD@MUC数据集上验证了所提出的使用字节对编码算法进行语音建模的BPE-Conformer-CTC/Attention模型效果，并与基线模型Transformer-CTC/Attention进行对比，如表2所示。

| 模型(CTC权重为0.3，beam size为6) | 建模单元 | XBMU-AMDO31 | | TIBMD@MUC | |
|---|---|---|---|---|---|
| | | 验证集 | 测试集 | 验证集 | 测试集 |
| Transformer-CTC/Attention | 构件 | 22.2 | 20.6 | 13.2 | 12.3 |
| | 字丁 | 16.3 | 15.2 | 9.8 | 9.7 |
| | 音节 | 11.2 | 10.5 | 6.8 | 6.6 |
| Conformer–CTC/Attention | 构件 | 9.7 | 9.5 | 5.5 | 5.1 |
| | 字丁 | 9.8 | 9.3 | 4.9 | 4.6 |
| | 音节 | 9.5 | 9.2 | 4.9 | 4.6 |
| BPE-Conformer-CTC/Attention | BPE500 | **9.4** | **9.0** | **4.7** | **4.6** |

**表 2.** 不同编解码器的WER(%)实验结果

实验结果表明，在TIBMD@MUC测试集上的BPE-Conformer-CTC/Attention模型效果与Conformer-CTC/Attention模型效果相当，相比于Transformer-CTC/Attention相对降低了30.3%。而在XBMU-AMDO31的测试集上，提出的BPE-Conformer-CTC/Attention模型相比于Conformer-CTC/Attention模型的词错误率相对降低了2.2%，相比于Transformer-CTC/Attention相对降低了14.2%。

CCL 2024

通过对比两个数据集，TIBMD@MUC文本长度较短，而XBMU-AMDO31文本长度较长且内容更加丰富，使XBMU-AMDO31数据集复杂度较高，从而突显了本文所提出的模型具有较强的鲁棒性。基于此，本文以下实验均在XBMU-AMDO31数据集上进行。

为了探讨BPE词典规模大小对藏语语音识别的影响，本文选择在XBMU-AMDO31数据集上设置不同的词典规模进行实验。实验结果如表3所示。

| BPE规模大小 | 模型 | XBMU-AMDO31 验证集 | XBMU-AMDO31 测试集 |
|---|---|---|---|
| BPE100 | ours | 9.9 | 9.6 |
| BPE300 | | 9.4 | 9.1 |
| BPE400 | | 9.5 | 9.1 |
| BPE500 | | **9.4** | **9.0** |
| BPE600 | | 9.7 | 9.2 |
| BPE800 | | 9.5 | 9.2 |

**表 3.** BPE规模对实验结果WER(%)的影响

选择适当大小的BPE词典对语音识别模型的性能至关重要。若选择过小的词典，会导致一些词汇被合并成单个子词，从而丢失语言的细微差别和语义信息，进而降低模型对文本的理解能力，尤其是处理语义丰富的文本时。而选择过大的词典则会增加识别困难度，使模型复杂度和过拟合风险增加。实验结果表明，当BPE词典大小为500时，能在词汇覆盖、模型性能之间取得平衡，提供良好的文本处理和理解能力。此时该模型的WER为9.0%。

为了更好的体现基于藏语文字构成规则与基于字节对编码算法进行语音建模时的差异，本文在对解码时搜索束大小beam_size进行讨论的同时，分别对比了以构件、字丁、音节和BPE500为语音建模单元的模型效果。对比结果如表4所示。

| 模型(CTC权重为0.3) | 建模单元 | XBMU-AMDO31 验证集 | XBMU-AMDO31 测试集 |
|---|---|---|---|
| beam_size= 2 | 构件 | 9.9 | 9.4 |
| | 字丁 | 9.8 | 9.7 |
| | 音节 | 9.8 | 9.4 |
| | BPE500 | **9.4** | **9.2** |
| beam_size= 4 | 构件 | 10.0 | 9.5 |
| | 字丁 | 9.8 | 9.6 |
| | 音节 | 9.6 | 9.1 |
| | BPE500 | **9.5** | **9.1** |
| beam_size= 6 | 构件 | 9.7 | 9.5 |
| | 字丁 | 9.8 | 9.3 |
| | 音节 | 9.5 | 9.2 |
| | BPE500 | **9.4** | **9.0** |
| beam_size= 10 | 构件 | 10.0 | 9.6 |
| | 字丁 | 9.9 | 9.5 |
| | 音节 | 9.3 | 9.2 |
| | BPE500 | **9.5** | **9.0** |

**表 4.** 探究beam-size大小以及建模单元的选择对实验结果WER(%)的影响

实验结果表明，纵向对比中，当beam_size的大小选为6时，四种建模单元的效果相较于其它beam_size大小的选择效果更好。横向对比中，相较于其它三种建模单元，BPE500作为建模单元效果最好。原因可能为以下几点：

（1）BPE通过反复合并出现频率最高的字符对，实现了对常见藏文字更好的捕捉。

（2）BPE能够减少稀疏性，特别是对于藏语中可能存在的稀有字符组合，能够有效提高模型的泛化能力。


(c) 2024 中国中文信息学会计算语言学专业委员会

（3）BPE能够更好地捕捉藏语的复杂语法结构，动态生成词汇，有助于提升模型在语音识别任务中的效果。

为了进一步验证混合CTC/Attention解码的有效性，本文通过改变CTC在联合优化损失中的占比进行实验。其中λ代表CTC所占的比重。CTC占比导致的模型性能如表5所示。

| 模型 | CTC占比 | XBMU-AMDO31 验证集 | XBMU-AMDO31 测试集 |
|---|---|---|---|
| BPE-Conformer-Attention | λ=0 | 11.6 | 11.1 |
| BPE-Conformer-CTC | λ=1 | 39.8 | 36.9 |
| BPE-Conformer-CTC/Attention | λ=0.2 | 9.6 | 9.0 |
| | λ=0.3 | **9.4** | **9.0** |
| | λ=0.5 | 9.5 | 9.3 |
| | λ=0.7 | 10.0 | 9.5 |

**表 5.** CTC占比对实验结果WER(%)的影响

实验结果表明，当CTC和Attention机制联合解码时，实验效果都优于CTC或Attention机制单独解码，将CTC损失函数占比设置为0.3时藏语语音识别的效果最佳，此时WER为9.0%，验证了混合CTC/Attention解码的有效性。

## 5 总结和展望

本文构建了BPE-Conformer-CTC/Attention端到端藏语语音识别模型，在公开的藏语语音数据集上探究了不同的建模单元以及联合参数对藏语语音识别的影响，验证了该模型对藏语语音识别的有效性。在未来工作中，将继续深入研究和完善本文中提出的语音识别方法，进而拓展到藏语多方言的语音识别研究领域，以不断提升藏语语音识别系统的性能和实用性。

## 参考文献


Amodei, Dario and Ananthanarayanan, Sundaram and Anubhai, Rishita and Bai, Jingliang and Zhu, Zhenyao. 2015. *Deep Speech 2: End-to-End Speech Recognition in English and Mandarin.* Computer Science.

Chorowski, Jan and Bahdanau, Dzmitry and Serdyuk, Dmitriy and Cho, Kyunghyun and Bengio, Yoshua. 2015. *Attention-Based Models for Speech Recognition.* Computer ence, 10(4):429–439.

Dong, Linhao and Xu, Shuang and Xu, Bo. 2018. *Speech-Transformer: A No-Recurrence Sequence-to-Sequence Model for Speech Recognition.* ICASSP 2018 - 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

Graves, A. 2006. *Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks.* Proc. Int. Conf. on Machine Learning, 2006.

Gulati, Anmol and Qin, James and Chiu, Chung Cheng and Parmar, Niki and Zhang, Yu and Yu, Jiahui and Han, Wei and Wang, Shibo and Zhang, Zhengdong and Wu, Yonghui. 2020. *Conformer: Convolution-augmented Transformer for Speech Recognition.*

高耀荣，边巴旺堆. 2023. 基于端到端深度学习的藏语语音识别研究. 现代计算机, 29(17):25–30.

Huang, Mingkun and Zhang, Jun and Cai, Meng and Zhang, Yang and Ma, Zejun. 2020. *Improving RNN transducer with normalized jointer network.*

黄晓辉，李京. 2018. 基于循环神经网络的藏语语音识别声学模型. 中文信息学报, 32(5):7.

Ko, Tom and Peddinti, Vijayaditya and Povey, Daniel and Khudanpur, Sanjeev. 2015. *Audio augmentation for speech recognition.* Interspeech 2015.

Lee, Jaesong and Watanabe, Shinji. 2021. *Intermediate Loss Regularization for CTC-based Speech Recognition.*



(c) 2024 中国中文信息学会计算语言学专业委员会

Li, Senyan and Li, Guanyu and Ning, Jiewen. 2022. *XBMU-AMDO31: An open source of Amdo Tibetan speech database and speech recognition baseline system.* National Conference on Man-Machine Speech Communication, NCMMSC2022.

Shibata, Yusuke and Kida, Takuya and Fukamachi, Shuichi and Takeda, Masayuki and Shinohara, Takeshi. 1999. *Byte Pair Encoding: A Text Compression Scheme That Accelerates Pattern Matching.*

Vaswani, Ashish and Shazeer, Noam and Parmar, Niki and Uszkoreit, Jakob and Jones, Llion and Gomez, Aidan N and Kaiser, Lukasz and Polosukhin, Illia. 2017. *Attention Is All You Need.* arXiv.

王松. 2019. 基于*LSTM-CTC*的藏语拉萨话语音识别系统. Master's thesis, 西北民族大学.

Yang, Xiaodong and Wang, Weizhe and Yang, Hongwu and Jiang, Jiaolong. 2020. *Simple Data Augmented Transformer End-To-End Tibetan Speech Recognition.* 2020 IEEE 3rd International Conference on Information Communication and Signal Processing (ICICSP).

Yue Zhao, Xiaona Xu, Jianjian Yue, Wei Song, Xiali Li, Licheng Wu, Qiang Ji. 2020. *An open speech resource for Tibetan multi-dialect and multitask recognition.* International Journal of Computational Science and Engineering.

CCL 2024

# 面向对话式阅读理解的高质量藏语数据集构建

达哇才仁[1,2]　朋毛才让[1,2]　孙媛[1,2,3,*]
[1]中央民族大学　信息工程学院，北京　100081
[2]国家语言资源监测与研究少数民族语言中心
[3]民族语言智能分析与安全治理教育部重点实验室
*通讯作者：孙媛
tracy.yuan.sun@gmail.com

## 摘要

对话式阅读理解作为对话式人工智能领域的重要研究方向，旨在使机器能够理解自然语言文本，并能够进行多轮对话以解答与文本相关的问题。随着生成式大模型的发展，该任务也成为评测大模型性能的重要指标之一。在此过程中，高质量数据集的构建成为该领域的关键任务。目前，相关算法模型在许多英语数据集上取得了显著进展，甚至超过了人类表现。然而，对于低资源语言，尤其是缺乏相应数据集的藏语，对话式阅读理解研究尚处于起步阶段。本文采用了一种人工与半自动结合的方法策略，构建了藏语对话式阅读理解数据集TiconvQA（Tibetan Conversational Question Answering）。该数据集共包含了20,358个对话对，涵盖了人物、地理和新闻三个领域。每一轮对话包括对话依据文本以及根据文本生成的多轮连续问答对。本文从对话数据的多样性、相关性、语言现象等方面对TiconvQA进行了详尽的分析与质量评估。并对藏文对话式阅读理解任务中存在影响评价指标的五种因素进行了优化。最终，我们采用了三种经典的对话式阅读理解模型以及藏文大模型TiLamb对数据集进行实验评估，实验结果验证了数据集的质量，并表明TiconvQA可用于模型在对话式阅读理解任务中的性能评测。

**关键词：** 藏文；对话式阅读理解；低资源语言；数据集

# Construction of high-quality Tibetan language dataset for conversational reading comprehension

**DawaCairen[1,2] PengmaoCairang[1,2] Yuan Sun[1,2,3,*]**
[1] School of Information Engineering, Minzu University of China, Beijing　100081
[2] National Language Resources Monitoring and Research Center for Minority Languages
[3] Key Laboratory of Ethnic Language Intelligent Analysis and Security Governance of MOE
*Corresponding author: Yuan Sun
tracy.yuan.sun@gmail.com

## Abstract

Conversational Reading Comprehension, as a significant research direction in conversational artificial intelligence, aims to enable machines to comprehend natural language texts and engage in multi-turn dialogues to answer questions related to the text. With the development of generative large language models (LLMs), this task has also become a critical benchmark for evaluating model performance. In this process, constructing high-quality datasets has become a key task in this field. Currently, related algorithm models have made significant progress on many English datasets, sometimes even surpassing human performance. However, for low-resource languages, especially Tibetan, which lacks corresponding datasets, research on conversational reading comprehension is still in its early stages. This paper adopts a combined strategy of manual and

第二十三届中国计算语言学大会论文集，第314页–第325页，太原，中国，2024年7月25日至28日。
卷1：主会论文
(c) 2024 中国中文信息学会计算语言学专业委员会

semi-automatic methods to construct the Tibetan conversational reading comprehension dataset, TiconvQA (Tibetan Conversational Question Answering). The dataset comprises 20,358 dialogue pairs covering three domains: personalities, geography, and news. Each dialogue includes the base text and multi-turn consecutive question-answer pairs generated based on the text. We provide a detailed analysis and quality assessment of TiconvQA, focusing on the diversity, relevance, and linguistic phenomena of the dialogue data. Additionally, we optimize five factors that influence evaluation metrics in Tibetan conversational reading comprehension tasks. Finally, we experimentally evaluate three classical conversational reading comprehension models and the Tibetan LLM TiLamb on the dataset. The experimental results validate the quality of the dataset and demonstrate that TiconvQA can be used for performance evaluation of models in conversational reading comprehension tasks.

**Keywords:** Tibetan , Conversational reading comprehension , Low-resource language , Dataset

## 1 引言

机器阅读理解（Machine Reading Compre-hension,MRC）是问答任务（Question Answer,QA）最受欢迎的变体之一，在自然语言处理社区引起了很多关注 (Gupta et al., 2020)。随着近年来大型语言模型的兴起，自然语言处理领域取得了前所未有的发展，针对主流语言的各种任务，大型模型的性能已经能够与人类相媲美 (Zhao et al., 2023)。然而，对于一些特定领域和低资源语言，尤其是在跨领域或跨语言情境下，这些模型的性能仍然不尽如人意。因此，在对话式阅读理解任务中仍然存在许多待解决的困难。

传统的单轮机器阅读理解任务无法实现连贯的对话回答。当面对问题中包含指代、推理、隐喻和省略等现实交流中常见语言现象的复杂情境时，传统模型往往难以做出准确的回答。为了更好地适应真实场景的需求，更复杂的挑战随之出现，即对话式机器阅读理解（Conversational Machine Reading Comprehension, CMRC） (Chen et al., 2023)。CMRC旨在使机器能够全面理解给定的文本段落，并能够对对话中的一系列问题做出恰当的回应，尽管其任务涉及到复杂的语言理解和推理，但其本质仍然是一项数据驱动的任务。因此高质量数据集的构建是其技术发展的关键基础。近年来，大规模数据集的引入和语言模型的涌现为CMRC任务的实现提供了可能性 (Lewis et al., 2019)。然而，现有的大多数CMRC数据集主要集中在中英文领域，如多轮对话数据集CoQA (Reddy et al., 2019b)和QuAC (Choi et al., 2018)，以及百度的中文问答数据集DuReader (He et al., 2017)等。这些数据集的推出推动了CMRC领域的研究。其中，以Reddy等人发布的大规模英文机器阅读理解数据集CoQA为代表，许多学者在该数据集上提出了自己的方法和模型。根据最新的官方数据，截至2024年3月发布的模型榜单中，在该数据集上最高得分为90.7分，超过人类水平1.9分，显示出显著的优势 (Reddy et al., 2019a)。

藏语作为中国少数民族语言之一，在国内外研究藏文信息处理的科研机构中，共享的数据资源十分有限。尽管近年来随着互联网的发展，网络上出现了大量的藏文信息，并且各种藏文信息处理相关工具也得到了良好的发展 (夏天赐and 孙媛, 2018) (龙从军et al., 2015) (安波and 龙从军, 2022)，但是藏文对话式阅读理解任务仍处于起步阶段。在这一任务中，最大的难题在于缺乏公开的高质量数据集，目前公开的藏文机器阅读理解数据集只有TibetanQA (Sun et al., 2021)。因此，构建一个面向对话式阅读理解任务的高质量藏文数据集成为推动藏文CMRC任务发展的迫切需求。

本文的主要贡献如下:

(1)本文设计了一个结合人工和半自动构建数据集的方法，并成功构建了一个对话式藏语机器阅读理解数据集TiconvQA，其中包含根据2,120个藏语文本段落构建的20,358个对话问答回合，内容涵盖了地理、人物和新闻等三个不同领域。数据集对话中的每一轮包括了一个问题和答案以及答案的相关证据文本。在构建过程中我们从文章收集、问题构建和校对审

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

核等方面均采用严格的流程，这包括确保对话轮之间存在关联性、对话自然性以及涵盖领域内容的多样性等特征。为了满足该领域相关研究人员的数据需求，我们已经将部分数据公开在https://github.com/TBNLP/TiconvQA。

(2)本文对数据集TiconvQA进行了详尽的分析比较，包括与经典数据集的段落长度、问题长度和对话轮数的平均值对比；数据集中问答类型的分布统计；以及通过抽样计算数据集中问答过程中各类语言现象的占比分布等多个角度进行了数据集质量的验证。此外，我们在实验过程中发现了评估该任务时存在的多个影响因素，包括答案的多样性、答案长度差异、藏语特有的书写特点以及分词不一致等方面。针对这些问题，我们提出了适用于该任务及其他藏语生成式任务评估的考量方法，并优化了评价指标的计算过程，以便为后续数据集评估提供参考。

(3)本文以三种经典的对话式阅读理解模型和藏语大模型Tilamb (Zhuang et al., 2024)作为TiconvQA数据集的基线模型，并展开了实验。同时我们进行了基于对话问答历史长度的消融实验观察对话问答历史长度对模型表现的影响，成功验证了数据集质量。最终，在经典的对话式阅读理解模型上的最高F1值达到了66.5%，藏语大模型Tilamb上的表现达到了72.8%。

## 2 相关工作

大规模数据集的发布是NLP（Natural Language Processing，NLP）领域近期快速发展的关键因素之一。这些数据集不仅为模型训练提供了必要的资源，而且推动了方法的创新和进步 (Torfi et al., 2020)。数据集的质量和规模直接影响到模型的理解能力和表现。对话式阅读理解研究的激增同样归功于大规模多轮对话数据集的出现。其中最为经典的包括英文数据集CoQA和QuAC，以及中文领域的DuReader数据集等。

CoQA数据集构建于涵盖7个不同领域的文本段落之上，总计包含8,000个对话，共计126,000个问答对。每个领域的数据均来源于特定的公开数据集，例如新闻领域的文章选自CNN/Daily Mail新闻数据集 (Hermann et al., 2015)。值得注意的是，Reddit文章和科学文章这两个领域的数据仅用于模型评估，并未参与模型训练，以此构成域外评估场景；而其余五个领域的数据则同时用于模型训练和评估，形成域内评估场景。CoQA数据集中的对话模拟了两个参与者之间的问答交互过程，其中一方提问，另一方根据文本内容进行回答。问题的形式不受限制，但要求答案能够在文本中找到充分的依据并体现一定的推理能力。虽然答案同样采用自由形式，并在文本中标注了相应的支持片段，但Yatskar等人 (Choi et al., 2018)发现，答案往往只是支持片段的轻微改写版本。CoQA数据集中的对话主要集中于细节信息的深入挖掘（约占所有问题的60%），而较少涉及诸如话题转换、澄清或定义等其他对话功能。评估指标则采用了宏观平均F1分数，并分别针对域内和域外场景进行计算，以全面衡量模型的性能。

QuAC数据集包含来自14,000个信息搜索对话框的100,000个问题。对话是根据维基百科文章中关于不同个体的部分准备的。数据集采用了不对称的设置：学生只能看到文章的标题和摘要，而老师可以看到整篇文章。这种设置迫使学生根据对话中获得的有限信息来探索隐藏问题的答案，而老师则通过提供文章的简短摘录（或在无法回答时回复“不知道”）来回答问题。由于数据集的不对称性，学生无法直接从文章中复制答案，这使得问题更具描述性、高度依赖上下文且开放性。与CoQA数据集相比，QuAC中的对话更频繁地切换到新的主题，从而增加了任务的复杂性。在该数据集中的答案是抽取式的，包括“是/否”、“不知道”以及从文章中提取的文本片段。此外，QuAC数据集不仅仅提供从文章中提取的答案片段，还包含额外的对话行为标注，如继续（跟进，可能跟进，或不跟进）。这些标注信号提供了额外有用的对话流信息，可用于训练当中，使模型能够更深入地理解对话的动态 (Zaib et al., 2022)。在评估过程中，除了整个数据集的宏观平均F1分数外，QuAC还评估了人类等效分数（HEQ） (Qiu et al., 2021)，衡量系统相对于普通人类的表现，通过比较系统的F1分数与人类的F1分数来确定系统是否达到或超过人类水平。

DuReader数据集是目前最大的中文机器阅读理解数据集，与前两者不同，中文的DuReader数据集则是一个大规模、面向真实应用、由人类生成的中文阅读理解数据集。其中所有的问题、原文都来源于实际数据（百度搜索引擎数据和百度知道问答社区），答案是由人类回答的。共包含了20万个问题、100万个文档和超过42万个人类总结的答案 (He et al., 2017)。

目前，英文和中文的对话式阅读理解数据集已经取得了显著进展，然而，在低资源语言领域的研究相对较少，这严重阻碍了低资源语言的对话式阅读理解的发展。在藏语领域，目前公


(c) 2024 中国中文信息学会计算语言学专业委员会

开的藏文机器阅读理解数据集仅有TibetanQA。TibetanQA包含了由人工手动根据1513篇文章构建的20000个问答对，形式为单轮机制，该数据集涵盖了自然、文化和教育等多个领域的文章 (Sun et al., 2021)。TibetanQA数据集的出现使得藏语机器阅读理解领域的尝试有了基础的数据支撑，然而，在藏语对话式阅读理解领域，数据集的空缺仍然是该领域发展的首要难点。为了解决这个问题，我们构建了首个藏语对话式阅读理解数据集，即TiconvQA。

## 3 构建过程

### 3.1 数据来源

我们从三个不同的领域中选择了文本段落作为TiconvQA数据来源：分别为来自云藏百科（https://baike.yongzin.com）的人物和地理词条信息，以及来自中国藏族网通（https://ti.tibet3.com）的新闻文章。在数据的清洗和整理工作中，我们对获取的原始数据进行了降噪处理，去除了图片、表格等非文本数据，最终得到了2427篇文章。然而，并非所有这些领域的文章都同样适合引发有趣的对话，因为只有一个实体的文章通常会导致问题完全集中在那个实体上。为了确保数据的质量和多样性，我们选取的文本长度被控制在200至600个音节之间。经过严格筛选，最终确定了2120条文本作为故事素材，用于构建对话问答。这些文本涵盖了丰富的主题和情节，有助于提高对话模型的应答能力和适应性。

### 3.2 构建方法

#### 3.2.1 人工构建

为了保证数据集的质量，我们选择组内三位研究机器阅读理解方向的藏族学生进行此数据集的构建工作。对于数据集的构建规则，我们将英文领域经典的多轮对话数据集CoQA作为主要基准，参考其构建思想和方案，保证构建过程的合理性和有效性。

| | |
|---|---|
| 标题 | སྒོག་དཔོན་ཙོན་མེ་གཙུ་རྫི་ན（吞弥 • 桑布扎） |
| 段落 | སྒོག་དཔོན་ཙོན་མེ་གཙུ་རྫི་ན་ནི་བོད་ཀྱི་ཡི་གེའི་སྐོལ་ཙོག་མཁར་གཤོད་པ་པོ་ཡིན། ཁོང་ནི་སྤྱི་ལོ་དྲུག་རབས་བདུན་པའི་ཤ་ཀ་མཉག་ར་ཀ་བར་སྐོན་པོ་ཙོན་མེ་ཨ་ནྲའི་སྐལ་ལ་འཁྲུངས།…… 吞弥 • 桑布扎是藏文的创制者，他出生在公元七世纪的 617 年，吞鲁热卡地方大臣吞弥 • 阿奴的家中。…… |
| 对话历史 | Q1：བོད་ཀྱི་ཡི་གེའི་སྐོལ་གཤོད་མཁན་ནི་སུ་ཡིན། 藏文的创造者是谁？<br>A1：སྒོག་དཔོན་ཙོན་མེ་གཙུ་རྫི་ན། 吞弥 • 桑布扎<br>R1：སྒོག་དཔོན་ཙོན་མེ་གཙུ་རྫི་ན་ནི་བོད་ཀྱི་ཡི་གེའི་སྐོལ་ཙོག་མཁར་གཤོད་པ་པོ་ཡིན། 吞弥 • 桑布扎是藏文的创制者<br>Q2：ཁོང་ནམ་དུས་སྐྱེ་འཁྲུངས་བ་རེད། 他何时出生？<br>A2：སྤྱི་ལོའི་དྲུག་རབས་བདུན་པའི་ཤ་ཀ་མཉག་ར公元七世纪的 617 年<br>R2：ཁོང་ནི་སྤྱི་ལོའི་དྲུག་རབས་བདུན་པའི་ཤ་ཀ་མཉག་ར་ཀ་བར་ 在公元七世纪的 617 年<br>Q3：ཡབ་ནི་སུ་ཡིན། 是谁的儿子？<br>A3：སྐོན་པོ་ཙོན་མེ་ཨ་ནྲ། 吞弥 • 阿奴<br>R3：སྐོན་པོ་ཙོན་མེ་ཨ་ནྲའི་སྐལ་ལ་འཁྲུངས། 大臣吞弥 • 阿奴的家中<br>…… |

图 1: 数据集构建示例

在人工构建多轮对话的过程中，由提问者和回答者两人协助完成。提问者根据文本提出问题，而回答者给出问题的回答以及答案证据。构建的数据集的示例如图1所示，其中展示了每个段落下的多轮问题与答案，展示了在多轮对话过程中，存在的指代和省略等多样的关联现象。此外，数据集中还包含了每个答案在原文中的出处，即证据。为确保数据集的质量，每个段落

第二十三届中国计算语言学大会论文集，第314页–第325页，太原，中国，2024年7月25日至28日。
卷1：主会论文
(c) 2024 中国中文信息学会计算语言学专业委员会

都构建了包含5-15轮问答对的对话。问题的类型涵盖了是非题、无法回答的问题以及直接从文本段落中获取答案的问题等。

### 3.2.2 半自动扩充

为了消除在数据集构建过程中的人类的不一致性和主观性，以及解决人工构建效率缓慢等问题，我们采用一种半自动的方式实现数据集的扩充。利用了大型语言模型在内容生成方面的先进能力，设计了一个基于ChatGPT3.5 (Achiam et al., 2023)的半自动数据集构建流程。随后，我们进行了人工的细致审核和校对，以确保数据集的质量和准确性。具体的构建流程如图2所示。

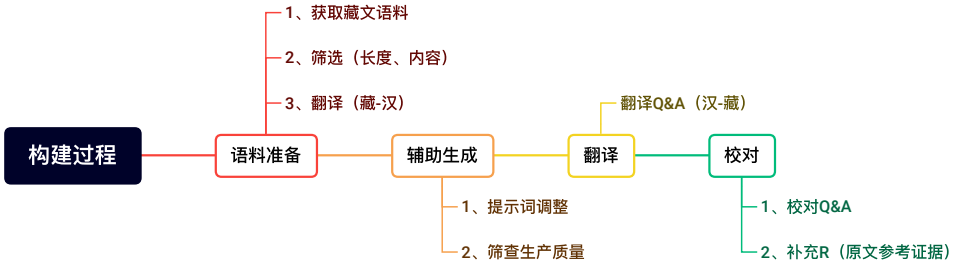



图 2: 半自动扩充构建流程

（1）预处理：首先，本文使用翻译工具将已准备好的文本转换为中文，这是因为当前最先进的大型模型主要集中在中英文等高资源语言上的研究，对于低资源语言（如藏文）的理解能力相对较弱。在这一阶段，关键在于不断调整和优化提示词，以确保生成的内容符合我们的质量标准和数据格式要求。

（2）自动化生成：完成预处理阶段后，调用ChatGPT3.5的API批量自动构建数据集。这一步骤充分利用了该模型在文本生成方面的优势，从而提高了数据集构建的效率和质量。

（3）人工审核和校对：在自动化生成多轮问答对之后，本研究通过人工进行细致的审核和校对，以确保数据集的准确性和可靠性。这一环节对于保证数据集质量至关重要，它有助于发现并纠正自动化生成过程中可能出现的错误和不一致之处。

在先前人工构建数据集的基础上，借助积累的构建经验，本文成功实施了半自动数据集构建流程，从而成功扩展了数据集规模。这种结合半自动和人工审核的方法不仅提高了数据集构建的效率，而且保证了数据集的质量。

## 3.3 构建难点与质量保证

在上述构建过程中伴随着许多挑战和难点。为了提升构建效率，同时注重数据集的质量，我们采取了一系列严格的方法和策略。以下是对这些挑战的介绍以及我们的应对措施：

（1）在人工构建过程中，常常存在主观臆断的情况，可能导致构建的数据集中的对话缺乏真实问答之间的连贯性。为了克服这一挑战，我们放弃了自问自答的构建方式，选择更加自然对答的方式，并在构建过程中随机地改变提问者和回答者的身份，通过这些方式以确保对话的构建是基于真实的交流模式和场景，尽可能模拟真实对话中的语言和逻辑交互。

（2）在人工反复构建数据集的过程中，存在陷入问答模式的风险。为了避免这种情况，我们设计了结合大型模型实现半自动扩充的方式，使得问答的过程更加自然随意。在此基础上，对话数据集的筛选和修订阶段经过严格的质量控制，以确保对话内容的合理性和准确性。通过这些措施，可以有效提高数据集构建效率和质量。

（3）数据集的校对和审核是非常关键。在完成结合半自动生成后，我们将最终结果翻译回藏文，并进行最后一轮的人工校对和质量审核。数据集最终的组成仍是以最原始的藏文文本及经过人工校对的对话内容。因此，需要对诸如翻译欠佳的问题进行修改、检查语法是否有问题、疑问词是否合理、表述是否自然流畅等质量把控。这种结合先进的自然语言处理技术和人工校对审核的方法，极大地提高了数据集构建的效率和质量。此外，这一流程还为未来其他低资源语言的数据集构建提供了一个可行的参考。

接下来，本文将着重介绍TiconvQA数据集的构建进展，并从数据集规模、问答类型、语言现象等多个角度分析我们的数据集，展示其在复杂性、多样性和对话相关性等方面的优势。

第二十三届中国计算语言学大会论文集，第314页-第325页，太原，中国，2024年7月25日至28日。
卷1：主会论文
(c) 2024 中国中文信息学会计算语言学专业委员会

# 4 数据分析

## 4.1 数据集规模

经过人工和半自动结合的构建策略，我们构建并扩充了TiconvQA数据集。最终，该数据集共包含来自2120篇文章的20358轮对话，具体构建情况如表1所示。

| 主题 | 构建方法 | 文章段落数 | QA对 |
|---|---|---|---|
| 人物 | 人工 | 669 | 5886 |
| 地理 | | 441 | 4304 |
| 新闻 | 半自动 | 1010 | 10168 |
| TiconvQA-总 | | 2120 | 20358 |

表 1: TiconvQA数据集及扩充部分构建情况

TiconvQA数据集的文本段落、问题和答案分别的平均字符长度为198、13.2和6.4，每篇文本的平均对话轮数为9.6轮。如表2所示，我们通过对比阅读理解任务的经典数据集SQuAD (Rajpurkar et al., 2016)，以及CoQA和QuAC发现，TiconvQA中的文本段落长度介于SQuAD和CoQA数据集之间，文本段落的长度直接影响着对话问答的轮数和难度。我们没有选择更长的文本段落，这是出于构建效率的考虑。考虑到团队规模的限制，过长的文本段落会明显降低数据集构建的速度。因此，文本的长度也间接影响了段落的平均对话轮数，TiconvQA相较于QuAC多出了约2.4轮。

| | TiconvQA | SQuAD | CoQA | QuAC |
|---|---|---|---|---|
| 段落平均长度 | 198 | 117 | 271 | 401 |
| 问题平均长度 | 13.2 | 10.1 | 5.5 | 6.5 |
| 答案平均长度 | 6.4 | 3.2 | 2.7 | 14.6 |
| 平均轮数 | 9.6 | - | 15.2 | 7.2 |

表 2: TiconvQA同经典数据集的对比

在对话式阅读理解任务中，对话分散在多个回合中，问题的平均长度侧面反映了对话中问答的关联性。在对话过程中，会存在许多省略、指代和非疑问句的连续问答，例如问题可以仅由一两个单词组成（谁？、何时？、为什么？），因此我们期望对话问题和答案比独立交互中的要短，从表中可以看出TiconvQA数据集的问题平均长度略高于CoQA和QuAC，答案平均长度则介于他们之间，在后续的构建中，我们将继续增加对话中问答的关联性。通过以上深入比较，能够初步的保证我们数据集的质量。

## 4.2 对话问答类型的分布

| 问答类型 | TiconvQA | SQuAD | CoQA |
|---|---|---|---|
| 可回答 | 93.1% | 66.7% | 98.7% |
| 不可回答 | 6.9% | 33.3% | 1.3% |
| 什么、哪个（名词短语） | 17.4% | 25.0% | 19.6% |
| 谁（命名实体） | 11.5% | 35.9% | 28.7% |
| 是否 | 9.7% | 0.1% | 19.8% |
| 多少（数字） | 23.9% | 16.5% | 9.8% |
| 何时（时间） | 18.4% | 7.1% | 3.9% |
| 其他 | 19.1% | 15.6% | 18.1% |

表 3: 问答类型统计对比

第二十三届中国计算语言学大会论文集，第314页-第325页，太原，中国，2024年7月25日至28日。
卷1：主会论文
(c) 2024 中国中文信息学会计算语言学专业委员会

在藏语中，每个问题类型都有着多样的复合形式。因此，在数据集构建的过程中，我们鼓励在提出问题时选择同义的其他复合疑问代词。此外，我们还模拟了实际对话中可能出现的无法回答问题和是非疑问句，以提高数据集的真实性和实用性。本文根据疑问词对提出的问题进行了统计分类。最终数据集中各问答类型的统计情况及与英文经典数据集的对比结果见表3所示。从各问答类型的分布情况来看，我们的数据集与经典数据集呈现出相似的趋势，各问答类型的比例相对均衡。值得注意的是，不可回答类型的比例高于SQuAD和CoQA数据集。

### 4.3 语言现象分析

在对话式阅读理解数据集的构建中，确保对话与文本段落以及对话历史之间的关联性，以及对话整体的连贯性至关重要。因此，我们进一步分析了对话与段落以及对话历史之间的关系。我们从数据集中抽样了150个对话问答对，并按照表4所示进行了各种现象的分析统计。

| 语言现象 | 示例 | 占比 |
|---|---|---|
| | 对话问题与段落之间的关系 | |
| 词汇匹配 | Q：ཁོ་སློབ་ཆེན་གང་ནས་**མཐར་ཕྱིན**་པ་རེད། 他**毕业于**哪所大学？ | |
| | A：བོད་ལྗོངས་སློབ་ཆེན། 西藏大学 | 29.33% |
| | R：བཀྲ་ཤིས་ནི་བོད་ལྗོངས་སློབ་ཆེན་གྱི་བོད་རིག་པའི་སྡེ་ཁག་ནས་**མཐར་ཕྱིན།** 扎西**毕业于**西藏大学藏学系 | |
| 释义重述 | Q：ཁོ་ལ་**བརྩམས་ཆོས**་གང་དག་ཡོད། 他的**作品**有？ | |
| | A：〔ལང་ཚོའི་རྒྱབ་ཆུ〕《青春瀑布》 | 49.33% |
| | R：ཁོང་གི་**སྙན་ངག**〔ལང་ཚོའི་རྒྱབ་ཆུ〕ཞེས་པ་ནི་བོད་ཀྱི་རང་མོས་སྙན་ངག་གི་ཐོག་མ་རེད། 他的**诗歌**《青春瀑布》是第一首藏文自由诗 | |
| 语用学 | Q：ཚེ་རིང་ནི་བུ་ཡིན་ནམ་བུ་མོ་ཡིན། 才仁是男生还是女生？ | |
| | A：མོ། **女** | 21.34% |
| | R：**ཁོ་མོ**་མཚོ་སྔོན་ཞིང་ཆེན་ཡུལ་ཤུལ་གྲོང་ཁྱེར་ནས་སྐྱེས། **她**出生在青海省玉树市 | |
| | 问题与其对话历史之间的关系 | |
| 无指代 | Q：**དགེ་ལུགས་པ**འི་སྲོལ་འབྱེད་མཁན་སུ་ཡིན། **格鲁派**的创教人是？ | 22% |
| 显性指代 | Q：གཉན་ཆེན་ཐང་ལྷ་རི་རྒྱུད་ཀྱི་རི་རྩེ་གཙོ་བོ་ནི་གང་ཡིན། 念青唐古拉山脉的主峰是？ | 48.67% |
| | A：རི་བོ་གཉན་ཆེན་ཐང་ལྷ། 念青唐古拉山 | |
| | Q：**དེའི**་མཐོ་ཚད་ག་ཚོད་ཡོད། **它**有多高？ | |
| 隐性指代 | Q：གསར་འགྱུར་ཁྱབ་བསྒྲགས་ཚོགས་འདུ་ནམ་དུས་འཚོགས་རྒྱུ་རེད། 新闻发布会何时召开？ | 29.33% |
| | A：ས་གནས་དེ་གའི་དུས་ཚོད་ཕྱི་དྲོའི་ཆུ་ཚོད་ཞ་དུས། 当地时间下午 3 点 | |
| | Q：**གང་དུ། 在哪里？** | |

表 4: 对话问题与段落之间的关系

在对话中问题与段落的关系中，我们将问题分类为词汇匹配，如果问题中至少包含一个在原文中出现的实词。这类问题约占总数的29.33%。若问题没有词汇匹配，但与原文释义相近，则归类为释义重述。这类问题涉及同义词、反义词、上下义和否定等现象，约占总数的49.33%。其余21.34%的问题没有词汇线索，我们将其分类为语用学，这些包括常识和预设等

CCL 2024

现象。例如表中的证据“她出生在青海省玉树市”并不与问题“才仁是男生还是女生？”存在直接相关，但是结合世界知识可以回答这个问题。对于问题与其对话历史的关系，我们将问题分类为是否依赖于对话历史。如果依赖于对话历史，则检查问题是否包含明确的标记。我们的分析表明，约22%的问题不依赖于与对话历史的共指，可以单独回答。将近一半的问题48.67%包含明确的共指标记，如他、她、那个等。这些指代对话中引入的实体或事件。其余的29.33%没有明确的指代标记，但隐含地指代了一个实体或事件。

## 5 藏文对话式阅读理解任务的评价指标优化

对话式阅读理解作为评估大型模型能力的重要任务之一，与其他NLP任务的不同之处在于其注重对给定文本的全面理解，并据此做出合理的回应。相比之下，情感分析下、命名实体识别和文本分类等任务更关注特定信息的提取和分类。因此，对话式阅读理解任务的评价过程应侧重于模型对文本理解和语境处理的能力，以及问题回答的准确性、完整性和语境相关性。通过观察对比实验中预测的答案和标准答案，我们发现在计算评价指标之前，存在着影响实验结果的五种因素，分别为答案的多样性、答案富裕或简洁的长度差异、藏语特有的书写特点以及分词不一致等。这些因素的具体实例如表5所示。为此，我们针对每一项进行了处理和改进，以优化后续的评价指标，从而减少评估过程中可能带来的误差，并全面评估对话式藏文机器阅读理解模型的性能。

| 影响因素 | 预测答案 | 标准答案 |
|---|---|---|
| 答案多样性 | ཏི17**སྣེ་ལེན་བྱས།**<br>**接待了** 17 万 | **མི་ཞེང་ས**ཏི17<br>17 万**人次** |
| 答案富裕 | **དཔྱིམས་སྣ་ཚོགས་འགྲུང་ཞིང་**ཐྱི་ངོས་འདབ་འབྲུར་མེ་སྐྱོམས་པ<br>**形状多样且**表面凹凸不平 | ཐྱི་ངོས་འདབ་འབྲུར་མེ་སྐྱོམས་པ<br>表面凹凸不平 |
| 答案简洁 | ལུང་རྟོགས་ཡོན་ཏན་འབྱམ<br>隆多元丁邦 | **སློབ་དཔོན**་ལུང་རྟོགས་ཡོན་ཏན་འབྱམ<br>隆多元丁邦**老师** |
| 书写特点 | ་གཞུང་ལུགས་དངོས་ལུགས་རིག་ར་<br>理论物理学 | གཞུང་ལུགས་དངོས་ལུགས་རིག་པ།<br>理论物理学。 |
| 分词不一致 | རྒྱལ་ཁབ་ །ཚན་རིག་ །ལག་རྩལ་ །ཡར་རྒྱས་ །བྱ་དགའ་<br>国家丨科学丨技术丨进步丨奖 | རྒྱལ་ཁབ་ །ཚན་རིག་ལག་རྩལ་ །ཡར་རྒྱས་བྱ་དགའ་<br>国家丨科学技术丨进步奖 |

表 5: 影响评价指标的因素与实例

（1）基于音节点的重切分：在本研究中，我们采用了藏语分词工具Tip-Las (李亚超et al., 2015)来进行数据的预处理。Tip-Las是一种基于条件随机场模型实现了基于音节标注的开源藏文分词系统，能够满足我们在实验中对藏语分词的需求。在最终进行评估指标计算的过程中，我们发现，即使答案是直接从原文中抽取的，当答案和原文分别进行分词时，仍会出现分词不一致的情况。这是因为分词工具会根据不同长度和内容的特征进行分词，导致不同的分词结果所致。因此，当我们在原文中预测答案的范围时，预测的答案可能与标准答案的分词情况不一致。如表中的例子所示，即使正确的答案也可能因分词不一致而导致评价指标计算出现误差，影响对模型性能的准确评估。为了解决这一问题，在最终的评估阶段之前，我们选择对句子进行重切分，将原有的分词结果进行合并。借助藏文中音节点的特性，我们对预测答案和标准答案都进行基于音节点的重切分，以消除分词不一致带来的影响，并在随后的评价指标计算中进行使用。

（2）答案的统一格式处理：藏语作为一种拼音语言，其最小单位为音节。通常情况下，音节之间通过特定标记音节点来进行分割，而单或双垂符则用于确定句子的停顿或结束。在中英文的评测过程中，常规做法是剔除符号和停用词等内容后进行计算。然而，在针对藏语的评估过程中，除了去除中英文的标点符号和停用词外，我们发现了一些其他情况可能会影响评价指

第二十三届中国计算语言学大会论文集，第314页-第325页，太原，中国，2024年7月25日至28日。
卷1：主会论文
(c) 2024 中国中文信息学会计算语言学专业委员会

标的计算。具体来说，在直接进行评价指标计算时，我们注意到音节点同时出现在词两端、音节点出现的位置不一致或缺失，以及其他藏语符号的出现等问题。如表5中的书写特点一栏所示。然而，这些情况不会影响答案的准确性，但会对评价指标的计算产生影响。为解决这类问题，我们选择对待评估的标准答案和预测答案进行统一的格式处理。通过基于音节点的重切分后，我们可以在评估计算过程中选择剔除所有音节点以及其他藏语符号和停用词，以减少对于评价指标计算的影响。

（3）EM指标的优化：在对话式阅读理解任务中，答案并非只有一个标准形式。在测试过程中，我们观察到模型预测的答案与标准答案可能在内容上存在富裕、简洁或表述不一致但任然合适的情况，如表5所示的实例。然而，传统的评价指标未能全面考虑到这种情况，当答案不完全匹配时，如EM评分会直接归零。这种过于绝对的评价指标无法有效评估模型的性能和数据集的质量，因此需要采用更为细致的评价方法，结合从语义层面考虑答案的相似性和正确性。为解决这一问题，我们对EM评价指标进行了优化称为$\mathrm{EM}_{semantic}$。该指标选择采用基于词向量的字符串相似度方法来衡量预测答案与标准答案之间的相似程度，并评估其指标。旨在更准确地衡量答案之间的语义相似性，而不仅仅局限于字符级别的匹配。

## 6 实验

对话式阅读理解数据集的质量直接关系到模型的理解能力。为了评估我们构建的多轮问答数据集TiconvQA的质量，本文分别采用了seq2seq (Gehring et al., 2017)、PGNet (Gu et al., 2016) 和DrQA (Chen et al., 2017)三种经典的机器阅读理解模型和藏语大模型TiLamb (Zhuang et al., 2024)上进行测试。其任务的核心是给定一段文本p，对话历史记录$\{ q_1, a_1, \ldots, q_{i-1}, a_{i-1}\}$和一个问题$q_i$，任务是预测答案$a_i$。

### 6.1 实验设置

本文将这四种模型作为基线方法，将数据集按照8∶2的比例随机划分为训练集和测试集,Tilamb模型中我们将训练集作为模型的任务微调数据。最后将采用宏平均的精确率、召回率和F1得分作为实验的评估指标。$\mathrm{EM}_{semantic}$值是针对该任务优化的EM指标，用于衡量模型是否在语义层面与参考答案相匹配，而F1值则综合考虑了精确率和召回率，用于度量模型生成答案与参考答案之间的相似度。举例来说，假设有n个对话问题，如果模型能够正确回答m*个对话问题（其中回答的准确性取决语义相似性的判断，而非传统的字符级别的完全匹配）因此，可以利用公式(1)来计算$\mathrm{EM}_{semantic}$值。

$$EM_{semantic} = \frac{m^*}{n} \tag{1}$$

F1值是准确率(precision)和召回率(recall)的调和平均，准确率，召回率和F1值的计算如公式(2)-(4)所示:

$$precision = \frac{N(TP)}{N(TP) + N(FP)} \tag{2}$$

$$recall = \frac{N(TP)}{N(TP) + N(FN)} \tag{3}$$

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{4}$$

其中，N（TP）表示预测的当前答案和标准答案之间相同的词数，N（FP）表示不在标准答案中而在预测答案中的词数，N（FN）是标准答案中的词而不是预测答案中的词数。

### 6.2 数据集在不同模型上的表现

Seq2Seq模型采用编码器-解码器结构，将输入序列编码为向量，再解码为目标序列。在对话式阅读理解中，它接受段落、对话历史和问题，生成回答序列。PGNet是Seq2Seq的扩展，添加了复制机制，可直接复制输入文本中的单词，提高了处理实体名词等未出现在训练数据中的情况的能力。DrQA基于阅读理解，包括阅读器和检索器，前者理解文本提供答案，后者从文档中找到可能包含答案的候选文档。在对话式阅读理解中，DrQA能有效处理段落和对话历


(c) 2024 中国中文信息学会计算语言学专业委员会

史，并回答问题。TiLamb不同于前三者，是一种基于开源基座模型LLaMA2-7B上进行增量预训练得到的藏文大语言模型。

| Model | TiconvQA | | CoQA | |
|---|---|---|---|---|
| | F1 (%) | $EM_{semantic}$ (%) | F1 (%) | EM (%) |
| 人类表现 | 89.5 | 80.2 | 88.8 | - |
| seq2seq | 23.9 | 13.0 | 20.9 | 17.7 |
| PGNet | 40.5 | 30.7 | 45.2 | 38.0 |
| DrQA | **66.5** | **44.5** | 55.6 | 46.2 |
| Tilamb | **72.8** | **45.3** | - | - |

表 6: TiconvQA在模型上的表现

我们将针对藏语的评价指标优化应用于实际评估中。最终，我们得到了四种模型在对话式阅读理解任务中的表现结果，见表6所示，我们观察到TiconvQA数据集的表现与CoQA数据集相近。特别是在DrQA模型上，其F1值甚至超过了CoQA数据集的表现，达到了66.5%。然而，在EM值方面却低于CoQA数据集。这一现象的主要原因在于，我们的数据集中的段落文本长度以及对话问答的轮数都低于CoQA数据集。因此，数据集的难度相对较低，还有待进一步提高。在Tilamb上的表现为最高，达到了72.8%。通过这次实验，我们更全面地了解了所构建的多轮问答数据集TiconvQA的质量，以及不同模型在此任务上的性能和优劣。这为未来的研究提供了重要参考，可以针对数据集难度的提高和模型性能的优化进行进一步的探索和改进。

### 6.3 基于对话问答历史长度的消融实验

不同对话历史长度的输入对结果的影响能够侧面反映对话式数据集的质量。为了进一步证明我们数据集的质量，我们进行了不同长度的对话问答历史信息与当前问题进行拼接的实验。我们分别考虑了将当前问题之前h轮的对话问答历史信息拼接到输入中的情况，其中h的取值包括h=0、h=2、h=4以及h=full四种情况。这里的h=0和h=full分别表示不拼接对话问答历史信息和将当前问题之前的所有对话问答历史信息都进行拼接。最终，我们得到了以下的实验结果。

| 对话历史长度 | Seq2seq | PGNet | DrQA |
|---|---|---|---|
| 0 | 21.92% | 35.95% | 61.51% |
| 2 | **23.92**% | 40.46% | **66.49**% |
| 4 | 21.65% | **41.49**% | 65.56% |
| all | 21.26% | 37.91% | 64.76% |

表 7: 不同对话问答历史长度对F1值的影响

采用不同长度的对话问答历史输入进行模型训练和预测时，模型的性能会发生变化。如表7所示，我们发现在所有模型上，这种变化的规律十分一致：当历史对话长度为0时，性能最差；当历史对话长度为2时，性能最佳；然而，当加入更多对话历史时，并不会提升模型性能，反而可能降低其性能。这是因为加入更多的对话历史可能会引入更多与问题无关的信息，从而影响模型的性能。通过这一实验结果，进一步验证了我们数据集的特性，包括问答的难度、对话间的相关性、对话的连贯性等方面，以及其高质量的体现。

## 7 总结与展望

本文以人工和半自动相结合的方式构建了藏文对话式阅读理解数据集TiconvQA，为藏文该领域的研究提供了数据基础。对该任务的评估指标进行分析与优化，目前，TiconvQA的基线模型达到了66.5%的F1值，在Tilamb上的表现达到了77.8%的F1值，进一步缩小了与人类表现的差距。同时，TiconvQA也为藏语大型模型提供了评估其对话式阅读理解能力的数据。未来，我们将持续扩展数据集的规模，并鼓励更多的研究者参与到新模型的探索中。通过不断改进模型


(c) 2024 中国中文信息学会计算语言学专业委员会

和优化扩展数据集，我们可以推动低资源语言对话式阅读理解技术的发展，并为解决相关问题提供更加有效的解决方案。

## 致谢

本论文得到了国家社科基金(22&ZD035)和国家自然科学基金项目（61972436），以及中央民族大学项目(GRSCP202316, 2023QNYL22, 2024GJYY43)的资助。

## 参考文献

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051*.

Nuo Chen, Hongguang Li, Junqing He, Yinan Bao, Xinshi Lin, Qi Yang, Jianfeng Liu, Ruyi Gan, Jiaxing Zhang, Baoyuan Wang, et al. 2023. Orca: A few-shot benchmark for chinese conversational machine reading comprehension. *arXiv preprint arXiv:2302.13619*.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. *arXiv preprint arXiv:1808.07036*.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *International conference on machine learning*, pages 1243–1252. PMLR.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor OK Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. *arXiv preprint arXiv:1603.06393*.

Somil Gupta, Bhanu Pratap Singh Rawat, and Hong Yu. 2020. Conversational machine comprehension: a literature review. *arXiv preprint arXiv:2006.00671*.

Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, et al. 2017. Dureader: a chinese machine reading comprehension dataset from real-world applications. *arXiv preprint arXiv:1711.05073*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Minghui Qiu, Xinjing Huang, Cen Chen, Feng Ji, Chen Qu, Wei Wei, Jun Huang, and Yin Zhang. 2021. Reinforced history backtracking for conversational question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13718–13726.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

S. Reddy, D. Chen, and C. D. Manning. 2019a. CoQA: A Conversational Question Answering Challenge. https://stanfordnlp.github.io/coqa/. Accessed on April 17, 2024.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019b. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Y Sun, S Liu, C Chen, Z Dan, and X Zhao. 2021. Construction of high-quality tibetan dataset for machine reading comprehension. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 208–218.

Amirsina Torfi, Rouzbeh A Shirvani, Yaser Keneshloo, Nader Tavaf, and Edward A Fox. 2020. Natural language processing advancements by deep learning: A survey. *arXiv preprint arXiv:2003.01200*.


(c) 2024 中国中文信息学会计算语言学专业委员会

Munazza Zaib, Wei Emma Zhang, Quan Z Sheng, Adnan Mahmood, and Yang Zhang. 2022. Conversational question answering: A survey. *Knowledge and Information Systems*, 64(12):3151–3195.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Wenhao Zhuang, Yuan Sun, and Xiaobing Zhao. 2024. Tilamb: 基于增量预训练的藏文大语言模型(tilamb: A tibetan large language model based on incremental pre-training). In *Proceedings of the 23th Chinese National Conference on Computational Linguistics*.

夏天赐and 孙媛. 2018. 基于联合模型的藏文实体关系抽取方法研究. 中文信息学报, 32(12):76–83.

安波and 龙从军. 2022. 基于预训练语言模型的藏文文本分类. 中文信息学报, 36(12):85–93.

李亚超, 江静, 加羊吉, and 于洪志. 2015. Tip-las: 一个开源的藏文分词词性标注系统. 中文信息学报, 29(6):203–207.

龙从军, 刘汇丹, 诺明花, and 吴健. 2015. 基于藏语字性标注的词性预测研究. 中文信息学报, 29(5):211–216.

CCL 2024

# 面向心理健康咨询的藏语数据集及大语言模型构建

**朱孟笑[1]　　沙九[2]　　冯冲[3*]**
1.北方工业大学信息学院
2.百度自然语言处理部
3.北京理工大学计算机学院
zhumx@ncut.edu.cn, shajiu_dn@163.com, fengchong@bit.edu.cn

## 摘要

焦虑、抑郁已成为人们常见的心理障碍，适度的疏导对于缓解人们精神、心理压力具有重要意义。然而由于病耻感等原因，很多人得不到及时的疏导和治疗。随着人工智能的发展，大语言模型（LLMs）优越的知识融会贯通能力和思维链能力，使得其成为心理疏导的有效工具。然而，现有少量面向心理健康咨询的大语言模型通常针对英文、中文等资源丰富的语种，而对于低资源语言，LLMs在心理咨询领域的应用尚缺少研究。本文以藏语作为低资源语言的代表，研究藏语心理咨询数据集的构建和藏语心理健康大语言模型的构建方法。首先，通过收集现有高质量的中文心理咨询对话数据，并对数据进行处理，生成心理健康多轮对话数据集；其次，构建汉藏翻译工具将其翻译成藏语多轮对话数据，并结合多种机制对数据进行筛选、过滤生成高质量藏语心理健康多轮对话数据；基于构造的数据，采用现有通用大语言模型Baichuan2和LLaMA2模型进行指令调优训练，形成藏语心理健康大语言模型，并将开源用于科学研究。最后通过实验验证了本文发布的藏语心理健康多轮对话数据集以及藏语心理健康咨询大语言模型的有效性。

**关键词：** 心理健康支持 ； 藏语 ； 大语言模型

# Construction of Tibetan Datasets and Large Language Models for Psychological Health Counseling

**Mengxiao Zhu[1]　　Shajiu[2]　　Chong Feng[3*]**
1. School of Information Science and Technology, North China University of Technology
2. NLP, Baidu Inc.
3. School of Computer Science and Technology, Beijing Institute of Technology
zhumx@ncut.edu.cn, shajiu_dn@163.com, fengchong@bit.edu.cn

## Abstract

Anxiety and depression have become prevalent psychological disorders, and moderate counselling plays a critical role in alleviating mental and psychological stress. However, due to reasons such as the sense of shame, many individuals do not receive timely counseling and treatment. With the advancement of artificial intelligence, large language models (LLMs) with their superior abilities in knowledge integration and cognitive chaining have become effective tools for psychological counseling. Nevertheless, existing psychological health LLMs are primarily focused on resource-rich languages like English and Chinese, with limited research on their application in low-resource languages. This paper focuses on Tibetan, a representative low-resource language, to explore the construction of Tibetan psychological counseling datasets and Tibetan psychological health LLMs. Initially, we collect high-quality Chinese psychological counseling dialogue data,

第二十三届中国计算语言学大会论文集，第326页–第339页，太原，中国，2024年7月25日至28日。
卷1：主会论文
(c) 2024 中国中文信息学会计算语言学专业委员会

process it, and create a multi-turn dialogue dataset for mental health; subsequently, we develop a Chinese-Tibetan translation tool to translate this into Tibetan, using multiple mechanisms to filter and produce high-quality Tibetan psychological health multi-turn dialogue data. Utilizing the constructed data, we fine-tune existing general LLMs, Baichuan2 and LLaMA2, to develop a Tibetan psychological health LLM, which will be open-sourced for scientific research. Finally, experiments validate the effectiveness of the released Tibetan psychological health multi-turn dialogue dataset and the Tibetan psychological health counseling LLM.

**Keywords:** Psychological health support , Tibetan , Large language model

## 1 引言

在当下，心理健康成为一个越来越重要的问题 (Liu et al., 2023a)。在快节奏和互联网时代，每个人都面临着很多影响心理健康的挑战。为了确保心理健康，及时的早期心理干预和诊断至关重要。然而，由于高昂的费用、复杂的语言程序和有限的资源，专业心理咨询师往往难以提供及时的援助。

大语言模型（Large Language Models，LLMs）的出现提供了一个可行的解决方案。随着ChatGPT、LLaMA等模型的发布，LLMs显示出了强大的意图理解、逻辑推理等能力，在很多任务上都取得了超出预期的效果。但现有大语言在面对心理咨询领域仍有一些局限性：（1）现有大语言模型是通用模型，缺乏心理健康领域的专业知识。心理健康咨询场景较为复杂，专业心理咨询师通常有丰富的领域知识，通用模型在面对心理咨询时，由于缺乏专业知识，往往出现避而不谈或答非所问的情况。（2）现有大语言模型在面对心理健康咨询时，通常急于解决用户的问题，给出冗长而笼统的建议，缺少追问能力。而专业的心理咨询师通常经过反复的提问来了解用户意图，给出专业、详实的回复。为了解决这些问题，少量研究构建了专门用于心理健康咨询领域的数据集，如SMILE数据集(Qiu et al., 2023)和AUGESC数据集(Zheng et al., 2022)等，还有个别工作采用GLM等模型作为基座模型，利用构建的数据集进行微调，实现心理健康领域的大语言模型 (Chen et al., 2023b)。

然而，现有心理健康咨询大语言模型通常针对英语、汉语等资源丰富的语种，依托丰富的开源数据可以构造高质量的心理健康咨询数据用于LLMs的指令微调。但这些LLMs难以很好地泛化到低资源语言中。对于文本数据充裕的语种，获取大规模数据样本相对容易。然而，在全球7000多种语言中，资源丰富的语言仅占少数，大部分语言的网络资源极为匮乏，许多低资源语言的网络数据占比不足0.1% (Liu, 2022)，这大大增加了收集这些语言数据的难度。此外，由于使用低资源语言的用户相对较少，对这些语言进行数据标注的工作也显得更为困难。标注数据的稀缺和语言间的显著差异意味着LLMs在低资源语言上的研究和应用仍处于起步阶段。因此，在低资源语言上，如何构造高质量的心理健康咨询数据、构建心理健康咨询大模型是一个亟待解决的问题。

本文以藏语作为低资源语言的代表进行研究。在我国，藏语是我国古老的语言文字之一，记载和传播了藏民族优良的文化和博大的精神，也在推动我国社会发展进步，民族团结一致等方面发挥了不可替代的重要作用。然而，藏语语言资源相对匮乏，在算法、算力和数据的平衡中，数据成为了制约藏语LLMs应用发展的关键短板。为推动心理健康大语言模型在藏语中的应用，需要构建藏族心理健康咨询数据集，进而构建藏语心理健康大模型。首先，为了解决数据稀缺的问题，选择中文高质量心理健康单轮对话数据集，并对数据进行自动化处理，生成中文高质量多轮对话数据集；然后，采用构建的汉藏翻译工具对数据集进行翻译，并采用多重筛选机制选择高质量的藏文心理健康多轮对话数据集；最后，基于现有开源的大语言模型，LLaMA-2和Baichuan-2，在构造的数据集上进行指令微调训练，构建藏语心理健康大模型[0]，通过分析大语言模型能力的提升程度验证数据集和模型的有效性。

本文主要贡献如下：

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版
基金项目：北方工业大学科研启动基金项目资助
[0]https://github.com/Shajiu/LLM/tree/main/Tibetan_Mental_Health_Chat

（1）收集高质量中文心理健康单轮对话数据集，基于GPT 4和藏汉翻译工具构造高质量藏语心理健康多轮对话数据集；

（2）基于多重筛选机制选择高质量藏文心理健康多轮对话数据集，并进行开源。

（3）选择LLaMA-2和Baichuan 2大语言模型作为基座模型，采用构建的藏文心理健康多轮对话数据集对基座模型进行指令微调，构建面向心理健康领域的藏语大语言模型MindL-Ti和MindB-Ti，并进行开源。

（4）通过实验分析本文构造的藏文心理健康多轮对话数据集的有效性，以及通过与基座模型进行对比，验证了藏语心理健康大语言模型的有效性。

## 2 相关研究

焦虑、抑郁等是世界范围内常见的心理障碍，互联网无处不在的特性催生了心理健康支持的新范式。随着大语言模型（LLMs）在各种领域中的广泛应用，其在心理健康领域的应用也正在逐渐展现出巨大的潜力。近期的研究工作表明，LLMs可以被有效地用于提供情感支持对话、精神健康咨询，并具有良好的共情和倾听技巧。而在这其中，高质量的心理健康数据对心理健康咨询模型的效果有着至关重要的作用。下面将对心理健康数据集构建和LLMs在心理健康支持领域的应用进行介绍。

### 2.1 心理健康数据集构建

近年来，关于心理健康支持的研究在很大程度上取决于公开数据集的可用性 (Qiu et al., 2023)。大规模的心理健康会话数据集对于识别心理健康状况 (Liu et al., 2023c; Srivastava et al., 2022)、决定个性化干预措施 (Golden et al., 2024)等具有重要的意义。

现有心理咨询领域的数据集通常是针对英文、中文等语种。在英文数据集中， Liu et al. (2021)通过众包的方式构建了情感支持会话数据集ESconv，包含1053个情感支持对话。由于该方法的收集时间和成本较高， Zheng et al. (2022)通过大语言模型来扩充ESConv数据集，首先基于ESConv数据集对GPT-J 6B模型进行微调获得从初始对话完成完整对话的能力；然后，收集多样化的对话帖子作为初始对话，基于微调后的模型生成完整对话，据此构建了包含102K个对话的AUGESC数据集；Crisis Text Line (Althoff et al., 2016)提供了一个高质量的心理健康咨询数据集，它包含由经验丰富的志愿咨询师进行的大规模多回合咨询对话，但这个数据集无法公开访问。在中文数据集中，心理咨询领域收个开放的问答语料库 (Hai Liang Wang, 2020)，包括20000条心理咨询数据，该语料库的内容由现实世界的用户提出，回答由志愿者给出，大多数回家比较简短且笼统。 Sun et al. (2021)在公开心理健康支持平台上爬取了对话帖子，构建了高质量的心理健康对话数据集PsyQA。 Qiu et al. (2023)则是将PsyQA中的单轮对话改写为多轮对话，增加了对话的丰富性和多样性。 Chen et al. (2023b)通过众包的方式收集了对话数据集，并基于ChatGPT进行重写，形成了超过200万的心理咨询多轮对话数据集。

### 2.2 LLMs在心理健康领域的应用

以ChatGPT为代表的大语言模型的出现重新定义了自然语言处理的范式，LLMs不仅显著超越了传统意义的人机对话系统，而且被视作以自然语言为交互方式的通用语言处理平台，具有超出预期的交互体验，包括通用的意图理解能力、强大的连续对话能力、较强的逻辑推理能力等 (Zhao et al., 2023; Wan et al., 2023)。随着LLMs的迅速发展，其应用范围越来越广，在文本生成 (Liebrenz et al., 2023)、医疗 (Jeblick et al., 2023)、机器人 (Vemprala et al., 2023)、编码辅助 (Liu et al., 2023b)等领域都有广泛的应用。

随着LLMs应用范围的扩大，以LLMs技术支撑的聊天机器人在心理健康咨询领域受到越来越多的关注。 Wang et al. (2023a)探讨了LLMs在提供心理健康咨询方面的应用。他们发现LLMs能够表现出一定程度的理解和同理心，提供的回答有助于用户的心理健康支持。 Qin et al. (2023)利用LLMs开发了一种可解释的交互式抑郁检测系统，为检测心理健康指标引入了一种新的范式； Chen et al. (2023a)探索了ChatGPT在模拟精神科医生和患者对话的潜力，该研究肯定了在精神科中部署ChatGPT驱动的聊天机器人的可行性，并深入研究了提示设计对机器人行为和用户参与度的影响。 Ayers et al. (2023)实现了一个基于ChatGPT的聊天机器人，比较医生和聊天机器人在社交媒体中对患者的回应，发现有78.6%的评估者认为聊天机器人更快、更善解人意。 Chen et al. (2023b)基于ChatGLM-6B作为基座模型，对心理健康领


（c）2024 中国中文信息学会计算语言学专业委员会

域的多轮对话数据集进行微调，构建支持心理健康咨询的模型SoulChat，使得模型的共情、倾听和安慰能力得到了显著提高。 Yang et al. (2024)基于大语言模型的多智能体系统，提出一种创新性的心理测量范式PsychoGAT，与传统问卷方式不同的是，该研究为每位参与者定制化生成一个可交互的叙事类型游戏，据此测量用户对应的心理特质。Mental-LLM (Xu et al., 2024) 分别基于Alpaca和FLAN-T5模型构建了针对心理健康领域微调的大语言模型Mental-Alpacha和Mental-FLAN-T5，并验证了模型的有效性。ChatCounselor (Liu et al., 2023a)是一个提供心理健康支持的大语言模型，与其他聊天机器人不同的是，其建立在真实对话基础上，使其拥有心理学领域的专业知识和咨询技能。该模型使用的训练数据集是由260次深度访谈构建而成，实验验证了高质量数据集对模型效果的改善。Psy-LLM (Lai et al., 2023)是基于盘古和闻仲大模型，采用心理健康咨询数据集对模型进行微调得到的心理健康支持的大语言模型。

这些最新的研究进展表明LLMs在心理健康领域的应用潜力正被逐步释放，这些方法为提供个性化的心理健康干预措施提供了新的途径。但这些模型都是基于丰富的开源数据集，因此现有关于心理健康支持的大语言模型通常是针对汉语、英语等资源丰富的语种，而对于低资源语种，由于开源心理健康领域数据集的匮乏，面向低资源语种的心理健康支持的大语言模型尚缺少研究。

## 3 面向心理健康支持的藏语数据集构建

### 3.1 数据来源

作为低资源语言，藏语数据极度稀缺，本文选择高质量的中文心理健康数据PsyQA (Sun et al., 2021)进行多重处理，生成藏语心理健康支持数据集。PsyQA数据收集自壹心理论坛，该论坛是一个心理健康服务平台，拥有着庞大的用户群体和专业的咨询师团队。在论坛中，匿名用户发布他们在日常生活中的烦恼和问题，由志愿者或专业咨询师给出文本回复。PsyQA包含了多种一般的心理健康障碍问题，设计了9个主题，包括自我成长、情感、爱情问题、人际关系、行为、家庭、治疗、婚姻和职业，其中，一个问题映射到多个答案。原始数据集统计如表1所示。

| 主题 | 问题数量 | 占比 | 答案数量 |
|---|---|---|---|
| 成长 | 4148 | 18.56% | 10585 |
| 情绪 | 3037 | 13.59% | 6804 |
| 恋爱 | 2956 | 13.23% | 8312 |
| 人际 | 2923 | 13.08% | 6911 |
| 行为 | 2490 | 11.14% | 5404 |
| 家庭 | 2466 | 11.04% | 6370 |
| 治疗 | 2304 | 10.31% | 5479 |
| 婚姻 | 1234 | 5.52% | 3962 |
| 职业 | 788 | 3.53% | 2236 |
| 汇总 | 22346 | 100% | 56063 |

Table 1: PsyQA原始数据

由于真实心理咨询场景中，咨询师往往通过多轮问询和解答来疏解用户的困惑，因此，单轮对话往往难以满足真实场景的需求，本文首先将其转换成多轮对话，然后将其翻译成藏语心理健康支持多轮对话数据，并对数据进行清洗等处理生成高质量的藏语心理健康支持多轮对话数据集。

### 3.2 多轮对话转换

由于原始数据集是论坛问答数据，为此在转换成多轮对话时，需要去掉一些在多轮对话中不会出现的内容，如“楼主”等 (Qiu et al., 2023)，此外，由于藏汉文化差异，需要删除独特于汉语的一些表达，如谚语等。本文基于规则的方法过滤掉相关内容，并人工进行检验与修正。

由于PsyQA中每个主题的数据量偏少，本文依据SELF-INSTRUCT方法 (Wang et al., 2023b)扩充单轮对话数据。将每个主题的单轮对话作为种子池，通过调用GPT-4生成单轮对


(c) 2024 中国中文信息学会计算语言学专业委员会

**转换之前：**

求助者：成绩不合格久久不能恢复，我怎么那么容易被挫折打倒？

支持者：楼主你好得知自己的成绩不合格，难受，虽然知道这个事情的主要原因是复习的少，但是还是忍不住的难受。楼主其实已经知道了原因，却还是很难受，那么楼主就可以去探究为什么难受？是因为自己没有看书么？是因为这次的题太简单？还是觉得自己没过会受到环境的影响？你需要把这里的关系理清楚，找到自己难受的根源，然后对症下药。楼主可以去听听音乐，做点运动，或者旅游放松自己的心情，其实楼主不必把这种事情放的太重，即便如此那又如何。你要问问自己，了解自己究竟想要什么，没有目标就不会用心对一件事情，有方向能坚持会事半功倍，加油。

**转换之后：**

求助者：成绩不合格，我久久不能恢复，我怎么那么容易被挫折打倒？

支持者：你好！面对成绩不合格确实很难受，大部分人都会有这样的感受。你已经意识到了复习不足可能是一个原因，这已经是一个很好的开始。你觉得，除了复习时间少，还有其他什么可能导致你这次考试没能达到预期吗？

求助者：可能是因为我没有很好地掌握学习方法，或者是我对这次考试的内容不够感兴趣。

支持者：这是很常见的情况。找到真正的原因可以帮助我们更好地解决问题。对于学习方法，你觉得有哪些方面是可以改进的？同时，对于兴趣的部分，有没有哪些科目或话题是你感觉更加投入的？

求助者：数学我通常比较感兴趣，但这次的章节我觉得特别难，所以就没那么积极。

支持者：那这就是一个很好的线索了。对于那些难的部分，我们可以尝试找一些不同的资源或方法来帮助你理解。同时，保持对自己感兴趣的部分的热情也很重要。你有尝试过和同学一起学习，或者向老师求助吗？

求助者：我试过几次，但感觉效果不是很好。

支持者：那也许我们可以尝试一些其他策略，比如寻找一些在线课程或视频，有时候不同的解释方式会有新的启发。此外，让自己在学习时保持放松也很重要，你有尝试过通过运动、听音乐或其他活动来放松吗？

求助者：有时候我会听音乐来放松。

支持者：那很好，保持心情的平衡对于学习效率是有帮助的。试着在学习和休息之间找到一个合适的平衡，不要太过压迫自己。你已经展现了很多自我反思的能力，我相信你能找到适合自己的学习方式，并逐渐改善成绩的。加油！

Figure 1: 多轮对话转换示例

话加入到原来的数据集中。然后，依据 Qiu et al. (2023)的方法，通过调用GPT-4将单轮对话扩展到多轮对话，采用的提示模版如表2所示。

| |
|---|
| 我想让你作为一个文本重写者。您应该遵循这些要求：1. 每个句子必须以求助者：或支持者为开始：；2.对话必须以求助者：为开始；3. 支持者的反应有适量的情感支持和调节；4. 每个说话人的话语长度应符合对话场景，不宜过长。<br>你需要将给定的单回合对话改写为心理健康支持对话，对话内容为求助者与支持者之间的10次或更多对话。<br>单论对话为：XXX<br>让我们一步一步地分析和重写它。<br>你重写的多回合对话是:”’ |

Table 2: 生成心理健康多轮对话的模版提示

图1通过一个例子展示了转换之前的单轮对话和转换之后的多轮对话，通过例子可以看出，以这种方式可以有效地将单轮对话转换成多轮对话。

### 3.3 多轮对话翻译

由于目前面向藏语的翻译工具比较稀缺，现有的翻译工具缺乏专门针对心理健康领域的翻译能力，尤其是在处理心理健康领域术语方面。为此，本文训练一个专门针对心理健康领域的汉藏翻译模型来对心理健康领域的中文数据进行翻译。然而，由于心理健康领域的汉藏平行语料非常稀缺，难以直接收集平行语料构建汉藏翻译模型。相对来说，获取心理健康领域的藏语单语语料比较容易，为此，本文采用反向翻译 (Sennrich et al., 2016)的方法来构建高质量的心理健康领域的汉藏翻译工具来获得高质量的藏文心理健康多轮对话数据集。


(c) 2024 中国中文信息学会计算语言学专业委员会

为了构建翻译模型，本文基于实验室收集的汉藏平行语料和CCMT 2020的藏汉翻译数据集，基于OpenNMT的Transformer模型训练翻译模型，构成一个基础的通用汉藏翻译模型。然后采用该模型，对收集到的心理健康领域的藏文单语数据集进行翻译，将其翻译成中文，称为“伪数据集”，该“伪数据集”基于翻译模型翻译得到，其中的中文会存在噪声，以此作为输入对翻译模型进行训练，可以提高翻译模型的泛化性。在得到“伪数据集”后，将其与原始的汉藏平行语料混合，形成更大规模的训练数据，用于训练最终的汉藏翻译模型。通过实验测试发现，通用汉藏翻译模型在心理健康咨询领域的BLUE值为41.86，而添加“伪数据集”后训练的汉藏翻译模型在心理健康咨询领域中多次测试的平均为43.39.

利用训练好的汉藏翻译模型对处理后的中文心理健康数据集进行翻译，获得藏文心理健康数据，并通过人工校正提升数据质量。

### 3.4 数据筛选

由于本文基于翻译的方式来获得藏文心理健康多轮对话数据集，翻译后的数据存在低质量的内容。本文结合多种机制对数据进行优化和筛选。

首先，基于N-gram语言模型对文本进行评估，过滤掉低质量的内容。本文将收集到的藏文文本数据分成N-gram序列，其中$N=3$，通过BERT-BASE训练藏语的$N$-gram语言模型。基于该模型对翻译后的藏文心理健康多轮对话中的每个语句进行$N$-gram序列的评分。针对每个多轮对话，其每个语句中所有$N$-gram序列评分的均值若低于某个阈值，则将其视为低质量内容。在本文中，通过实验分析发现阈值过大时，容易过滤掉大量正常文本，而阈值过小时，低质量内容无法被过滤掉，平衡过滤的效果，将阈值设置为0.85.

其次，基于指令跟随难度（Instruction-Following Difficulty, IFD）(Li et al., 2023)对多轮对话数据进行筛选，提高模型指令调优的有效性。而由于指令跟随难度需要根据大语言模型的回答进行评估，而现有通用大语言模型在藏语心理健康咨询领域缺乏指令跟随能力，为此，本文采用少量高质量的数据对模型进行指令调优训练，进而利用调优后的模型来评估数据的指令跟随难度。由于本文收集的数据包含9个主题，为了保证质量和多样性，从每个主题中人工挑选等量的数据，构成1000条高质量数据集。在获得数据集后，利用该数据集对LLaMA-2模型进行指令调优训练，获得预经验模型。接着利用该模型对翻译后的心理健康多轮对话数据进行预测，通过指令内容预测答案，并获取预测答案与真实答案之间的差异，采用条件回答分数（Conditioned Answer Score, CAS）来衡量该差异，CAS为调优过程中的平均交叉熵损失。在调优过程中，给定问答对$(Q,A)$，其损失计算为：

$$L_\theta(A|Q)=\frac{1}{N}\sum_{i=1}^{N}\log P(w_i^A|Q,w_1^A,...,w_{i-1}^A;\theta) \tag{1}$$

其中，$\theta$为模型的参数，$N$为回答$A$中单词的数量，$w_i^A$表示回答$A$序列中的第$i$个词，$P(w_i^A|Q,w_1^A,...,w_{i-1}^A;\theta)$是在问题$Q$和回答序列$w_1^A,...,w_{i-1}^A$的条件下得到词$w_i^A$的概率。CAS值的大小反应了模型基于指令生成答案的难易程度，但其可能受到模型生成答案$A$难易程度的影响。为此，采用直接答案分数（Direct Answer Score, DAS）来衡量模型生成答案的难易程度：

$$s_\theta(A)=\frac{1}{N}\sum_{i=1}^{N}\log P(w_i^A|w_1^A,...,w_{i-1}^A;\theta) \tag{2}$$

其中，$P(w_i^A|w_1^A,...,w_{i-1}^A;\theta)$是在回答序列$w_1^A,...,w_{i-1}^A$的条件下得到词$w_i^A$的概率。DAS得分越高，表示模型生成该答案任务本身更为复杂。为了获取更好的指令数据，需要去除答案本身的影响，因此采用IFD分数来评价指令数据对模型调优的影响，计算如下：

$$IFD=\frac{CAS}{DAS} \tag{3}$$

较高的IFD分数表示模型难以将答案与给定的指令内容进行对齐，表示指令的难度更高，对模型调优更为有利。通过计算每个问答对的IFD指标，选择分数更高的数据作为藏语心理健康多轮对话数据集，最终数据集如表3所示。


(c) 2024 中国中文信息学会计算语言学专业委员会

| 主题 | 规模（条数） | 占比 |
|---|---|---|
| 成长 | 9296 | 18.592% |
| 情绪 | 8074 | 16.148% |
| 恋爱 | 6912 | 13.824% |
| 人际 | 6846 | 13.69% |
| 行为 | 4980 | 9.96% |
| 家庭 | 4932 | 9.86% |
| 治疗 | 4608 | 9.22% |
| 婚姻 | 2468 | 4.94% |
| 职业 | 1884 | 3.77% |
| 汇总 | 50000 | 100% |

Table 3: 藏语心理健康多轮对话数据集分布表

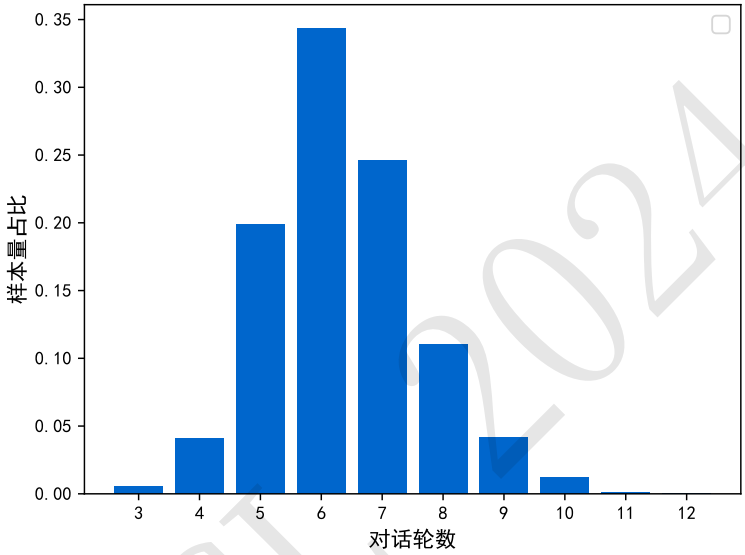



Figure 2: 藏语心理健康多轮对话轮数分布

基于以上方式最终形成了50000个藏语心理健康多轮对话数据，每个多轮对话的轮数在3-12轮之间，轮数的分布如图2。图 3是其中一个多轮对话示例。对最终的藏文心理健康多轮对话数据进行统计，结果如表4所示。其中，平均长度是去除“求助者”和“支持者”之后的文本按照音节计算出来的长度，这里的音节通过藏语文本中的”་”进行分割得到。

| 对话数目 | 平均轮数 | 平均长度 | 平均长度（求助者） | 平均长度（支持者） |
|---|---|---|---|---|
| 50000 | 6.34926 | 57.493 | 41.298 | 73.688 |

Table 4: 数据统计信息

## 4 藏文心理健康咨询大语言模型构建

### 4.1 基座模型选择

本文选择两个通用大语言模型LLaMA-2和Baichuan-2作为基座模型，其中，LLaMA-2（Large Language Model Meta AI） (Touvron et al., 2023)是由Meta AI发布的一个开放且高效的大型基础语言模型，作为完全开源的模型，被广泛用作AI模型的通用基座；Baichuan-2 (Yang et al., 2023)是国内百川智能发布的一个开源大语言模型，多次在国内外测评榜单中脱

CCL 2024

Q: ང་རང་ཉེ་ལམ་མགོ་འགོར་ཡོང་ཐམོ་གཞུམ་རྩོན་ཀྱི་ང་རང་མཐོ་རྒྱུགས་རྩོན་ཀྱི་འཇིགས་སྐག་ལས་གཞན་དུ་དུང་རང་ཉིད་སྐྱེས་མ་ཐག་ནས་བཟུང་སྐག་མ་ཞིག་ཡིན་པར་འདོད་དོན་སྙིང་ཅི་ཡང་མེད་པ་ལ་སོགས་པོ་མེད།ཁེར་རྒྱང་བཙོང་མི་ཐུབ།ཁྱོད་ཀྱིས་ང་ལ་བསམ་འཚར་འགའ་འདོན་ཐུབ་བམ། (我最近有点迷茫，高三之后的我除了高考前的恐惧，还觉得自己生来就是多余的，没有什么意义，我没有朋友，孤独难耐，你能给我一些建议吗？)
A: ཁྱོད་ཀྱི་སྣང་ཚ་མཚོང་རེས།ཁྱོད་ལ་སེམས།ཁྱལ་ཆེན་པོ་སྐྱེས་བྱང་།འདི་ནི་རྒྱན་མཚོང་གི་མཐོ་འཛིང་འཚོ་བའི་རྣམ་པ་རེད།ཐོག་ཁྱོད་ཀྱི་ཚོར་སེམས་མཚོང་པོ་ཞིག་ལ་མེད་པ་རེད།མཐོ་རྒྱུགས་ཀྱི་ཆེས་གལ་ཆེ་ན་ཁོ་སེམས།འགམས་རེད།ཁྱོད་ལ་རྒྱལ་པ་ཐོབ་པའི་སེམས།འགམས་དགོས་པ་རེད། (看到你的话，我很担心你，这是常见的高中生活状态，首先想说的是，我看到了你的决心，这是很多人都没有的，高考最重要的是心态，你要有成功的心态。)
Q: ཁྱོད་ཀྱིས་སྐུལ་མ་བྱས་པར་བཀའ་དྲིན་ཆེར་དངོས་གནས་སྐོང་གཡེར་ཐུང་འདོད།འོན་ཀང་ཁྱོད་ཀྱི་དམིགས་འབེན་དང་སྐུལ་ཤུགས་ཤོར་མི་རྩང་བར་འདོད། (谢谢你的鼓励，我真的很想放松，但我觉得自己不能失去目标和动力。)
A: གལ་ཏེ་ཁྱོད་ཀྱིས་འཚར་གཞི་ཅི་རྟར་བཟོ་དགོས་པ་མི་ཤེས་ན།ངས་བསམ་འཚར་འགའ་འདོན་ཐུབ། (如果你不知道如何制定计划，我可以提些建议。)
Q: འོ་ན་ངས་འཚར་གཞི་ཅི་རྟར་བཟོ་དགོས་སམ། (那我应该怎么制定计划呢？)
A: མཐོ་གཅིག་ནས་འགོ་བརྩམས་ཏེ་གོམ་པ་གང་ནས་གང་རེ་སྟོས་ན།འགོ་རྩོམ་གཅིག་ལས་ཐེབ་ཙམ་ཡོད་ཀྱི་ཐོག་ནས་ཁྱོད་ཀྱིས་ཐུབ་ཀྱི་ཡོད་ཚུལ་ཡ་རྒྱང་གཞིའི་གཏན་ཁེལ་ཡོད་ཚུལ་ཁྱོད་ཀྱིས་ཤེས་ཐུབ།

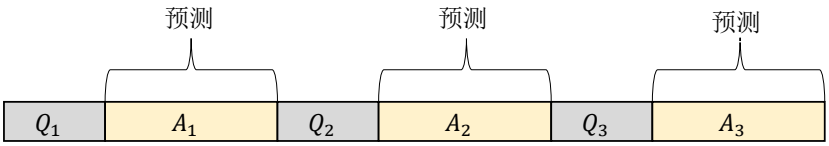



Figure 4: 藏文心理健康多轮对话数据训练方法示意图

Token_id

| $< s >$ | $Q_1$ | $</s >$ | $A_1$ | $</s >$ | $Q_2$ | $</s >$ | $A_2$ | $</s >$ | $Q_3$ | $</s >$ | $A_3$ | $</s >$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|

A_mask

| 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|

Figure 5: 掩码向量示意图

据$Q_1, A_1, Q_2, A_2, Q_3, A_3$，将该多轮对话数据进行拼接后输入模型，并计算每个位置的经验损失，只有$A_1, A_2, A_3$部分的经验损失参与模型参数的更新。由于以GPT为代表的因果语言模型都采用了一种注意力掩码机制，即对角掩码矩阵，在该机制中，每个token在编码时，只能看到它之前的token信息，而看不到它之后的token信息，因此在该三轮对话中，$Q_1$部分的编码输出只考虑了$Q_1$的内容，可以用来预测$A_1$的内容；而$Q_2$部分的编码输出是考虑了$Q_1, A_1, Q_2$的内容，可以用来预测$A_2$的内容；$Q_3$部分的编码输出是考虑了$Q_1, A_1, Q_2, A_2, Q_3$的内容，可以用来预测$A_3$的内容。对于整个序列，只需要输入一次就可以获得每个位置的经验损失，从而更新模型参数。

为了采用这种方法进行训练，需要选择属于因果语言模型的大语言模型。因此本文选择的是LLaMA2-7B和Baichuan2-7B作为基线模型进行训练。在训练时，为每个回复后面添加结束标记符，如$< /s >$作为此轮对话生成结束的标识符。因此，将$n$轮的多轮对话拼接成如下格式作为模型的输入：

$$< s > Q_1 < /s > A_1 < /s > Q_2 < /s > A_2 < /s > ... \tag{4}$$

此外，为了标记文本中的token属于提问还是回复，生成一个掩码向量$A_mask$，取值为0或1，用来标记文本中的token是否属于回复$A$部分，即是否需要模型进行预测，如图 5，其中$A < /s >$部分的掩码全为1，其余部分均为0.

根据这种方式，并行计算每个位置的经验损失，只有掩码$A_mask$=1的部分位置的损失函数才会参与模型参数更新。该方法充分利用了模型并行计算的优势，并且多轮对话中每个回复$A$部分都参与了训练，提高了数据的利用率。

本文采用Firefly (Yang, 2023)作为大语言模型指令调优训练的工具，选择QLORA (Dettmers et al., 2024)作为指令微调技术，使用本文的指令微调数据集分别对基座模型LLaMA2-7B和Baichuan2-7B进行指令微调，得到藏文心理健康咨询大语言模型MindL-Ti和MindB-Ti。

## 5 实验

### 5.1 实验实现

对基座模型LLaMA2-7B和Baichuan2-7B模型进行指令调优训练的大部分相关参数设置遵从Firefly的默认设置。由于采用QLoRA进行训练，QLoRA矩阵的秩$lora_rank$设置为64，QLoRA中的缩放参数$lora_alpha$设置为16，训练时的最大长度$max_length$设置为1024，学习率$learning_rate$设置为2e-4，训练步数$trainingstep$设置为1500。本实验的模型都是在拥有4块32G NVIDIA Tesla V100的服务器上进行调优训练与测试。

为了验证构造的藏文心理健康多轮对话数据集及藏文心理健康咨询大语言模型的有效性，采用定量与定性分析结合的方法进行。为了方便评估，本文开发了一个在线藏语心理健康咨询平台，该平台让专业咨询师能够向每位客户提供免费的文本形式咨询服务，平台如图6所示。




Figure 6: 在线藏语心理健康咨询平台

定量分析采用BLEU-4、ROUGE-1、GOUGE-2、GOUGE-L作为评价指标，来测试本文构造的数据集对于大语言模型能力的提升效果，以及构造的大语言模型MindL-Ti和MindB-Ti相比基座模型在心理健康咨询领域效果的提升。本文采用5折交叉验证的方式进行实验，将本文构造的藏文心理咨询多轮对话数据集随机分成5份，各占比20%，采用其中四份作为训练集进行训练，另外一份作为测试集。取5次测试结果的平均值作为最终结果进行展示。

## 5.2 定量分析

| 模型 | ROUGE-1 | ROUGE-2 | ROUGE-L | BLUE-4 |
|---|---|---|---|---|
| Baichuan2-7B-base | 55.07 | 28.55 | 21.64 | 2.94 |
| Baichuan2-7B-chat | 48.57 | 28.34 | 19.78 | 8.14 |
| Llama2-7B-base | 40.55 | 23.99 | 27.76 | 10.96 |
| Llama2-7B-chat | 25.99 | 15.63 | 7.38 | 5.76 |
| MindB-Ti | 83.14 | 41.54 | 44.58 | 28.27 |
| MindL-Ti | 65.04 | 34.14 | 40.42 | 27.36 |

Table 5: 各模型在测试集上的实验结果

对于定量分析，利用本文构造的心理健康多轮对话数据集进行指令调优训练的模型MindB-Ti和MindL-Ti与对应的基座模型Baichuan2-7B-base、Baichuan2-7B-chat、LLaMA2-7B-base和LLaMA2-7B-chat在测试集上的实验结果如表 5所示。根据表中的结果可以看出，本文构建的藏文心理健康多轮对话数据集对于提升大语言模型的能力具有显著的效果，通过在两个基座模型上进行调优训练，得到的模型MindB-Ti和MindL-Ti相比基座模型在各指标上都有较大的提升，在（ROUGE-1、ROUGE-2、GOUGE-L、BLUE-4）上相比base模型提升的平均值分别为（28.07，12.99，22.94, 25.33）、（24.49, 10.15, 12.66, 16.4），相比chat模型提升的平均值分别为（34.57，13.2，24.84, 20.13）、（39.05, 18.51, 33.04, 21.6）.


(c) 2024 中国中文信息学会计算语言学专业委员会

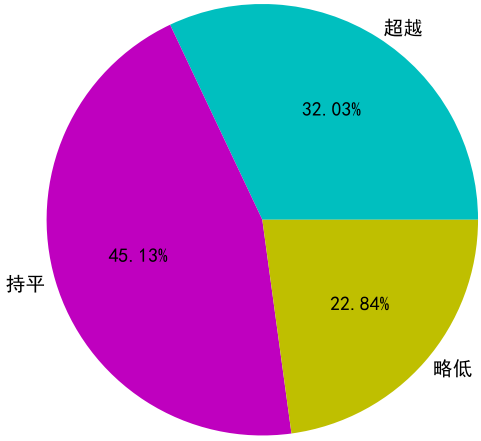



Figure 7: 人工评估结果

Q: ངའི་ཕྱིས་པ་སློབ་གྲྭར་འགྲོ་བའི་སྐབས་གུ་རྒྱན་དུ་དཔལ་འགོར་ཁེད་ལ་དང་།ཁོ་སློབ་གྲྭར་སྐྱེལ་ཞིང་ཤ་ཚོད་ཟེལ་འཚུབ་སྐྱེལ་གུ་བཏུག་དས་ཀྱང་ཁོ་ལ་སྐད་ཆ་ཙམ་འདོད་ཐིབ།་འདི་ཅི་ཤར་ཁེད་ལ་ཐག་ཆོད་ཡོག་ཡིན།་འདི་ཅི་ཤར་ཁེད་དགོས་སམ། (我的孩子上学时总是磨蹭，每次送他上学都让我很着急，我也会对他大喊大叫，这应该怎么办呢？)

A: དས་ཁྱོད་ཀྱི་ཚོར་བ་ཤེས།ཁྱོད་ཀྱིས་བཤད་པ་ཤར་ན།ཁྱོད་ཀྱི་ཕྱིས་པར་རང་ཉིད་ཀྱི་ཚོར་བ་སྐྱེན་པོ་དང་རྒྱུན་ལྡན་རེད།འབའ་ཞིག་ཏུ་ཁྱོད་ཀྱི་ཚོར་བ་ཙམ་མིན་པའི་ཁེད་ཀྱི་ཐབས་ལམ་ཡང་སྟོང་ཁྱེད་དགོས་པ་རེད། （我知道你的感受，据你说，你的孩子有自己的敏感和正常，这是很多家长的想法，你不能仅表达自己的感受，你要注意自己的情绪。）

Q: ཨེན།་ངས་བསམ་ན་ཁྱོས་པ་དང་ཁྱེམ་བདག་ཀ་ཤས་ཀྱི་ཐེམས་ཁམས་རྒྱུ་རྐྱེན་ཁྱོས་གནད་དོན་འདི་དག་བཟོས་པ་རེད།ཏ་ལའི་དག་ཡིན་ན་ངས་ཅི་ཤར་ཁང་ཁེད་ཀྱི་ཐེམས་ཁམས་ལ་ཚོད་འཛིན་ཁྱེད་དགོས། (嗯，我想是有一些孩子和家长的心理因素导致了这些问题。那我该如何控制自己的情绪呢？)

A: ཐོག་མར།ཁྱོད་ཀྱིས་རང་གི་ཐེམས་ཁམས་ལ་རྒྱས་པོན་ཁྱེད་པ་དང་ཉེ་ནས་ཐབས་བསོར།འཚེན་ནས་ཐག་གཅོད་ཁྱེད་པར་ཚོད་རྟེན་ཅིག་ཁྱོད་ཀྱིས་དཔྱག་རྩལ་ཁྱེད་ཚོགས་ཤ་ཟན་ཡིན་ན་རང་ཉིད་ཀྱི་ཐེམས་ཁམས་བརྩོལ་ཏོག་ཙམ་ལ་ཉེན་ཁྱེད་ཀྱི་ཐབས་ཤེས་ཡོད་པ་ཨིན་ན།ཁྱོད་ཀྱིས་ཐེམས་ཁམས་ལ་ཚོད་འཛིན་ཁྱེད་མི་ཐུབ་པར་འདོད་ན།ཁེད་ཀྱིས་ཐེམས་ཁམས་ལ་ཚོད་འཛིན་ཁྱེད་ཚོགས། （首先，你可以试着了解你的情绪，然后想办法解决，比如，你可以深呼吸、听音乐或者表达自己的情绪，如果你觉得情绪失控，可以找一些专业人士咨询。）

Q: བོད་ཀྱང་ངས་ཁྱོས་པར་ང་གཞའ་ནན་མེན་པར་འདོད་ནས་ང་ལ་བརྡེ་བཀུར་མི་ཁྱེད་པར་སེམས་ཁྱོད། (但是我也担心孩子会觉得我不够严格，不尊重我。)

A: ངས་ཁྱོད་ཀྱི་སེམས་ཁྱོ་ཤེས།ཁོས་ཀྱང་ཁྱོས་ཀྱིས་ཁྱོས་པ་ཚོར་ཁྱོད་ཀྱི་སེམས་ཁྱོ་བཤད་ཚོགས།ཁྱོས་ཀྱི་སེམས་ཁམས་བརྩེ་ཚོགས།ཁྱོས་པ་འདི་དག་ལ་སེམས་ཁམས་བརྩེ་ཚོགས།ཁྱོས་པ་ཚོར་ཁྱོད་ཀྱི་སེམས་ཁམས་བརྩེ་ན།ཁོ་ཚོས་ཁྱོད་ཀྱི་སེམས་ཁམས་ལ་རྒྱུས་པོན་ཁྱེད་ཤུལ་པ་མ་རེད།ཁྱོད་ཀྱི་སེམས་ཁམས་ལའང་རོགས་རམ་དང་རྒྱབ་སྐྱོར་ཁྱེད་ཤུལ། （我知道你的关注点，但是你可以告诉孩子们你的意思，你可以告诉孩子们你的心情，你在关心这些孩子，告诉他们你的心情，他们不仅能理解你的情绪，也能帮助和支持你的情绪。）

Figure 8: 采用MindL-Ti模型的案例

### 5.3 定性分析

对于定性分析，为了评估本文提出的藏语心理健康咨询大语言模型的性能，本文采用专业人员评估的方式进行。首先，从测试集中随机抽取了100个多轮会话，每个会话包含至少9轮对话。接着，采用基座模型和本文构造的藏语心理健康咨询大语言模型中生成了相应的回应。然后，将历史对话、由基线模型生成的回应、本文构造的模型生成的回复和真实回复一起提供给专业心理咨询师，由他们根据回应的自然度、类人特性、正确性等方面选择出历史对话的最佳回应。本文采用多数投票原则，由5位具有心理健康咨询背景的人员给出最终决策，结果如图 7所示。根据结果显示，采用本文构造的心理健康咨询大语言模型展现出了良好的性能。在所有回复中，约有32.03%回复在某种程度上超出了真实回应；达到真实回应相当水平的回应约占45.13%，而约有22.84%的回复略低于真实回复的水平。

此外，为了直观展示本文提出的藏语心理健康咨询大语言模型的有效性，选取了一个心理健康咨询的案例进行展示，使用相同的问题作为模型的输入，分析不同模型对问题的回答是否合理。由于未使用本文构造的藏语心理健康多轮对话数据微调的模型往往给出冗长的回复，本文仅展示LLama2-7B-chat模型针对不同问题的回复情况，如图9所示。对于微调后的模型，采用与之对应的MindL-Ti模型的结果进行展示，结果如图8所示。根据图中的结果可以看出，未使用本文构建的藏语心理健康多轮对话数据进行微调的模型在回答藏语心理健康问题时，往往给出无意义的回复。例如，在回答第一个问题时，Llama2-7B-chat模型对问题进行了复述，并重复了很多无意义的短句；在回答第二个和第三个问题时，直接用英语分析了藏语句子，无法直接对藏语问题进行直接回答。而使用本文构建的藏语心理健康多轮对话数据进行微调的模


(c) 2024 中国中文信息学会计算语言学专业委员会

Q:ངའི་ཉེས་པ་སྐོབ་གྲྭར་འགྲོ་བའི་སྐབས་ལུ་རྒྱུན་དུ་དལ་འགོར་ཁེར་ལ་དང་།ཁོ་སྐོབ་གྲྭར་སྐྱེལ་ཞིང་རར་ང་ཤ་ཙང་ཟེལ་འཚུབ་སྐྱེལ་ལུ་བསྟག་ངས་ཀྱང་ཁོ་ལ་སྐད་ཆེན་པོས་འབོད་སྟིང་འདི་ཅི་ཕྱར་ཉེར་དགོས་སམ།
A:ངའི་ཉེས་པ་སྐོབ་གྲྭར་འགྲོ་བའི་སྐབས་ལུ་རྒྱུན་དུ་དལ་འགོར་ཁེར་ལ་དང་།ཁོ་སྐོབ་གྲྭར་སྐྱེལ་ཞིང་རར་ང་ཤ་ཙང་ཟེལ་འཚུབ་སྐྱེལ་ལུ་བསྟག་ངས་ཀྱང་ཁོ་ལ་སྐད་ཆེན་པོས་འབོད་སྟིང་འདི་ཅི་ཕྱར་ཉེར་དགོས་སམ།ཡོངས་ཁྱོད་དགོས་སམ།ཡོངས་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།དོན་མི་ཁྱོད་དགོས་སམ།ཨེ（我的孩子上学总是磨蹭，每次送他上学都让我很紧张，我也会对他大喊大叫，这该怎么办呢？该办吗？该办吗？不该办？不该办吧？ 到底该不该当？该不该当？该当吗？该当事吗？该当事吗？该当吗？该当么？不当回事吗。该当么？不当吧？ 到底该不该当？该不该当？该当当？该打人还是该打人？该不该打事？该不该打事？该做事？不该做事？不该当事吗？该当事吗？应该做人吗？ 到底该不该当？该不该当？该当吗？该当事吗？该当事吗？该当吗？该当么？不当回事吗。该当么？不当吧？ 到底该不该做，到底该不该做，到底是不是该做。要不要？事要做？事要做？人要做事吗？事需要做吗？事需要做吗？不需要做什么吗？不需要做什么吗？ 到底该不该做，事不该做，事不该做。）
Q:ཨེན།ངས་བསམ་ན་ཉིམ་པ་དང་ཐྱིམ་བདག་ཀ་ཤས་ཀྱི་ཚེམས་ཁམས་རྒྱ་ཆེན་ཐྱིས་གཟད་ཅོན་འདི་དག་བཟོས་པ་རེད།ན་ངས་ཁོ་ལ་ཟེར་ཀྱི་ཚེམས་ཁམས་ལ་ཚོང་འཛིན་ཉིད་དགོས།
A:This is a Tibetan text that contains a list of 108 different types ofatial beings or "lower realms" that are said to exist in Buddhist teachings. These beings are believed to inhabit various planes of existence, including hells, hungry ghosts, animals, and humans, and are said to be trapped in a cycle of suffering and rebirth due to their negative actions in past lives. The text provides detailed descriptions of each of the 108 types of beings, including their physical characteristics, their suffering, and the causes of their suffering. The text also includes advice on how to help these beings through the practice of Buddhism and the performance of rituals and offerings.
Q: བོན་ཀུང་ངས་ཀྱང་ཉེས་པས་ང་ཁ་ཟས་འདོད་ནས་ང་ལ་བརྒྱབ་ཁཀུར་མི་ཉེད་པར་སེམས་ཁྱོད་ཀྱི།
A:This is a Tibetan sentence, which can be translated to English as: "The one who has the power to do something, but does not do it, is like a person who has a treasure but does not use it.“ In this sentence, the word "ངས" (ngas) means "power" or "ability", and "ངྱང་ཉེས་པས" (ngyang bgyis pa) means "to have the power to do something". The word "ང་ཁ་ཟས་ནན་མི" (ngag zhabs ngan mi) means "a treasure", and "རྩེ་བཀུར་མི་ཉེད་པར" (rtsis bkuur mi byed pa) means "to use it". The sentence is using a metaphor to compare the person who has the power to do something but does not do it, to someone who has a treasure but does not use it.

Figure 9: 采用Llama2-7B-chat模型的案例

型MindL-Ti在回答藏语心理健康问题时，能够给出比较合理且人性化的回复。例如，在回答第一个和第三个问题时，该模型首先肯定了求助者的感受，然后给出了合理的建议。这样的回复方式更符合心理健康咨询场景的问答方式，有助于舒缓求助者的心情，并且更容易让求助者接受合理的建议。

## 6 结论

基于大语言模型的心理健康咨询对个性化的心理健康干预具有重要意义。针对现有心理健康咨询大语言模型难以泛化到低资源语言的问题，本文以藏文作为低资源语言的代表，基于GPT 4和藏汉翻译工具构造高质量的藏文心理健康多轮对话数据集，并基于多重筛选机制选择高质量的数据构成最终的藏文心理健康多轮对话数据；然后，利用该数据构造藏文心理健康大语言模型，选择LLaMA 2和Baichuan 2大语言模型作为基座模型，利用构建的心理健康咨询多轮对话数据对基座模型进行指令微调，构建了面向心理健康咨询的大语言模型MindL-Ti和MindB-Ti。最终的实验验证了本文构造的藏文心理健康多轮对话数据集和藏文心理健康大语言模型的有效性。

然而，由于藏文具有特殊的语法语义结构，现有大语言模型中的原始词汇表中缺少藏文词汇，虽然一些大语言模型，如LLaMA模型通过回退到字节来支持其他语言的字符，但这种回退策略增加了序列长度，并且难以很好地体现语言的语义信息，因此如何针对藏文数据构造词表，进行二次预训练对于提升大语言模型在藏文上的理解和生成能力具有重要意义。

## 参考文献

Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics*, 4:463–476.

John W Ayers, Adam Poliak, Mark Dredze, Eric C Leas, Zechariah Zhu, Jessica B Kelley, Dennis J Faix, Aaron M Goodman, Christopher A Longhurst, Michael Hogarth, et al. 2023. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA internal medicine*, 183(6):589–596.

Siyuan Chen, Mengyue Wu, Kenny Q Zhu, Kunyao Lan, Zhiling Zhang, and Lyuchun Cui. 2023a. Llm-empowered chatbots for psychiatrist and patient simulation: application and evaluation. *arXiv preprint arXiv:2305.13614*.

Yirong Chen, Xiaofen Xing, Jingkai Lin, Huimin Zheng, Zhenyu Wang, Qi Liu, and Xiangmin Xu. 2023b. Soulchat: Improving llms' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1170–1183.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Grace Golden, Christina Popescu, Sonia Israel, Kelly Perlman, Caitrin Armstrong, Robert Fratila, Myriam Tanguay-Sela, and David Benrimoh. 2024. Applying artificial intelligence to clinical decision support in mental health: What have we learned? *Health Policy and Technology*, page 100844.

Jia Yuan Lang Hai Liang Wang, Zhi Zhi Wu. 2020. 派特心理：心理咨询问答语料库.

Katharina Jeblick, Balthasar Schachtner, Jakob Dexl, Andreas Mittermeier, Anna Theresa Stüber, Johanna Topalis, Tobias Weber, Philipp Wesp, Bastian Oliver Sabel, Jens Ricke, et al. 2023. Chatgpt makes medicine easy to swallow: an exploratory case study on simplified radiology reports. *European radiology*, pages 1–9.

Tin Lai, Yukun Shi, Zicong Du, Jiajie Wu, Ken Fu, Yichao Dou, and Ziqi Wang. 2023. Psy-llm: Scaling up global mental health psychological services with ai-based large language models. *arXiv preprint arXiv:2307.11991*.

Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. 2023. From quantity to quality: Boosting llm performance with self-guided data selection for instruction tuning. *arXiv preprint arXiv:2308.12032*.

Michael Liebrenz, Roman Schleifer, Anna Buadze, Dinesh Bhugra, and Alexander Smith. 2023. Generating scholarly content with chatgpt: ethical challenges for medical publishing. *The lancet digital health*, 5(3):e105–e106.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. *arXiv preprint arXiv:2106.01144*.

June M Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi Liao, and Jiamin Wu. 2023a. Chatcounselor: A large language models for mental health support. *arXiv preprint arXiv:2309.15461*.

Yue Liu, Thanh Le-Cong, Ratnadira Widyasari, Chakkrit Tantithamthavorn, Li Li, Xuan-Bach D Le, and David Lo. 2023b. Refining chatgpt-generated code: Characterizing and mitigating code quality issues. *ACM Transactions on Software Engineering and Methodology*.

Yuhan Liu, Anna Fang, Glen Moriarty, Robert Kraut, and Haiyi Zhu. 2023c. Agent-based simulation for online mental health matching. *arXiv preprint arXiv:2303.11272*.

Zihan Liu. 2022. *Effective Transfer Learning for Low-Resource Natural Language Understanding*. Hong Kong University of Science and Technology (Hong Kong).

Wei Qin, Zetong Chen, Lei Wang, Yunshi Lan, Weijieying Ren, and Richang Hong. 2023. Read, diagnose and chat: Towards explainable and interactive llms-augmented depression detection in social media. *arXiv preprint arXiv:2305.05138*.


(c) 2024 中国中文信息学会计算语言学专业委员会

Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. 2023. Smile: Single-turn to multi-turn inclusive language expansion via chatgpt for mental health support. *arXiv preprint arXiv:2305.00450*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In Katrin Erk and Noah A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 86–96.

Aseem Srivastava, Tharun Suresh, Sarah P Lord, Md Shad Akhtar, and Tanmoy Chakraborty. 2022. Counseling summarization using mental health knowledge guided utterance filtering. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3920–3930.

Hao Sun, Zhenru Lin, Chujie Zheng, Siyang Liu, and Minlie Huang. 2021. Psyqa: A chinese dataset for generating long counseling text for mental health support. *arXiv preprint arXiv:2106.01702*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Sai Vemprala, Rogerio Bonatti, Arthur Bucker, and Ashish Kapoor. 2023. Chatgpt for robotics: Design principles and model abilities. *arXiv preprint arXiv:2306.17582*.

Zhongwei Wan, Xin Wang, Che Liu, Samiul Alam, Yu Zheng, Zhongnan Qu, Shen Yan, Yi Zhu, Quanlu Zhang, Mosharaf Chowdhury, et al. 2023. Efficient large language models: A survey. *arXiv preprint arXiv:2312.03863*, 1.

Jiaqi Wang, Enze Shi, Sigang Yu, Zihao Wu, Chong Ma, Haixing Dai, Qiushi Yang, Yanqing Kang, Jinru Wu, Huawen Hu, et al. 2023a. Prompt engineering for healthcare: Methodologies and applications. *arXiv preprint arXiv:2304.14670*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023b. Self-instruct: Aligning language models with self-generated instructions. In *ACL*, pages 13484–13508.

Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K Dey, and Dakuo Wang. 2024. Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1):1–32.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Qisen Yang, Zekun Wang, Honghui Chen, Shenzhi Wang, Yifan Pu, Xin Gao, Wenhao Huang, Shiji Song, and Gao Huang. 2024. Llm agents for psychology: A study on gamified assessments. *arXiv preprint arXiv:2402.12326*.

Jianxin Yang. 2023. Firefly(流萤): 中文对话式大语言模型. https://github.com/yangjianxin1/Firefly.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Chujie Zheng, Sahand Sabour, Jiaxin Wen, and Minlie Huang. 2022. Augesc: Large-scale data augmentation for emotional support conversation with pre-trained language models. *arXiv preprint arXiv:2202.13047*.

CCL 2024

# 融合多元特征表示的藏文命名实体识别方法

**俄见才让**[1,2]　**周毛克**[2,3]　**陈波**[1,2]　**赵小兵***[1,2]
[1]中央民族大学　信息工程学院　北京　100081
[2]国家语言资源监测与研究少数民族语言中心　100081
[3]中央民族大学　中国少数民族语言文学学院　北京　100081
*通讯作者：赵小兵
{nmzxb_cn@163.com}

## 摘要

本文针对基于音节嵌入方式的藏文命名实体识别(TNER)中词汇信息和音节部件信息忽略的问题，提出了基于交叉Transformer架构的MECT-TL模型，融合了藏文音节信息、词汇信息和音节部件信息的多元数据特征。MECT-TL通过平面网络结构将藏文音节与词汇信息结合，并整合音节部件信息，有效提升了藏文实体识别的准确性。实验结果显示，相较于主流的TNER基准模型BiLSTM-CRF，本文模型在F1值上提高了5.14个百分点，与基于Transformer架构的TENER模型相比提高了4.18个百分点。这表明，融合藏文词汇和音节部件信息的方法可以显著提高TNER任务的性能。

**关键词：** 词汇信息；音节部件信息；多元数据特征；交叉Transformer

# Research on Tibetan Named Entity Recognition Using Multi-Feature Fusion Representation

**Cairang Ejian**[1,2]　**Maoke Zhou**[2,3]　**Bo Chen**[1,2]　**Xiaobing Zhao***[1,2]
[1]School of Information Engineering, Minzu University of China, Beijing,100081
[2]National Language Resources Monitoring and Research Center for Minority Languages,100081
[3]School of Chinese Ethnic Minority Languages and Literatures, Minzu University of China, Beijing,100081
*Corresponding author: Xiaobing Zhao
{nmzxb_cn@163.com}

## Abstract

This paper addresses the issue of neglecting lexical and syllable component information in Tibetan Named Entity Recognition (TNER) based on syllable embedding. It proposes the MECT-TL model based on a cross-Transformer architecture, which integrates multiple data features including Tibetan syllable information, lexical information, and syllable component information. MECT-TL combines Tibetan syllables with lexical information through a flat network structure and integrates syllable component information, effectively improving Tibetan entity recognition accuracy. Experimental results show that compared to the mainstream TNER baseline model BiLSTM-CRF, the proposed model achieves a 5.14 percentage point improvement in F1 score, and a 4.18 percentage point improvement compared to the Transformer-based TENER model. This indicates that integrating Tibetan lexical and syllable component information significantly enhances the performance of TNER tasks.

**Keywords:** Lexical information , Syllable component information , Multiple data features , Cross-Transformer

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

## 1 引言

命名实体识别(NER)在信息理解会议(MUC-6)(Grishman and Sundheim, 1996)首次提出以来，迅速成为自然语言处理领域的核心研究课题。从文本中自动识别并分类具有特定意义的实体，对于深入理解和分析不同语言和文化背景下的文本至关重要。NER在信息检索 (Khalid et al., 2008)、知识图谱构建 (Riedel et al., 2013)、问答系统 (Diefenbach et al., 2018)、舆情分析 (Wang et al., 2016)、生物医学研究 (Settles, 2004)、推荐系统 (Karatay and Karagoz, 2015)和机器翻译 (Babych and Hartley, 2003)等多个下游任务中发挥关键作用。

藏文命名实体识别(TNER)由于其独特文化价值和语言特性，对于推动藏文智能处理和文化传承具有重要意义。在实现TNER任务时，通常采用基于藏文音节的方法。与基于词的方法相比，这种方法通过实验证明通常更优，因为它不依赖于分词，从而避免了分词错误带来的影响。然而，这种只关注单个字符的方法不仅无法充分捕捉词汇的语义信息，还忽略了藏文音节内部结构信息。

TNER模型进行训练或评估时，词汇在预测实体边界起到至关重要的作用，例如，输入数据通常是已经标注好的序列文本如图1(a)所示，输入藏文例句:“མཚོ་སྔོན་པོ་ཏྲ་ཚན་རྩལ་ཀུང་སི་”被翻译为“青海布达科技公司”，这句话中包含了地理实体:“མཚོ་སྔོན་(青海)” 和组织实体“པོ་ཏྲ་ཚན་རྩལ་ཀུང་སི་(布达科技公司)”。数据集文本序列使用BMES标注方案标注相应的实体类型标签，以帮助模型学习如何识别和分类文本中的不同实体。如图1(b)所示，基于藏文音节的藏文命名实体识别方法因处理每个单独的藏文字符，从而缺乏对整体词义和词边界的认识。导致模型在实体边界的预测上不够准确，错误地将图1(a)的输入文本序列中的实体识别为地理实体：“མཚོ་སྔོན་པོ་(青海湖)”，非实体：“ཏྲ་(达)”，组织实体：“ཚན་རྩལ་ཀུང་སི་(科技公司)”。这种误判揭示了基于藏文音节方法在捕捉词义和辨认词边界上的不足，倾向于产生不准确的实体边界预测。直接影响到实体边界的准确预测，导致识别准确率下降，尤其开放领域的文本处理中，这一问题尤为突出，因为文本的多样性和复杂性更高，分词错误更难避免，其对实体识别的影响也更为显著。词汇信息在确定实体边界中扮演着至关重要的角色，如图1(c)所示，通过在藏文音节级别嵌入中融合藏文词汇信息，模型能够更准确地识别和划分实体。将输入序列文本正确识别为地理实体“མཚོ་སྔོན་(青海)”，组织实体:“པོ་ཏྲ་ཚན་རྩལ་ཀུང་སི་(布达科技公司)”,与原来标注数据集标签完全对应。这种在藏文音节嵌入基础上结合词汇嵌入的方式可以融入更多文本特征信息以判定词汇边界不仅能够显著提高TNER的准确性和鲁棒性，还能增强模型对文本结构的理解能力。




(a) 输入文本序列




(b) 藏文音节嵌入




(c) 藏文音节融合藏文词汇嵌入

图 1. 藏文词汇信息对TNER的影响

每一个藏文音节都是由精细的音节部件构成，每一部件都遵循藏文的语法规则，共同形成具有特定含义的藏文音节。每个部件都承载着特定的语义和功能。一个完整的藏文音节由一个基字或者基字基础上1至6个音节部件构成，这些部件包括前加字、上加字、下加字、元音、后加字以及重后加字。例如，图2所示“བསྒྲིགས(排列)”这一藏文音节，就由上述六种部件组成。

藏文音节部件不仅在视觉上具有区分度，而且在语言处理中扮演着独特的角色，它们对于区分词汇的含义至关重要。因此，基于藏文音节嵌入的方法忽视这些音节内部的特征，等同于忽略了藏文本身的一些关键信息，这可能会导致模型在识别过程中


(c) 2024 中国中文信息学会计算语言学专业委员会

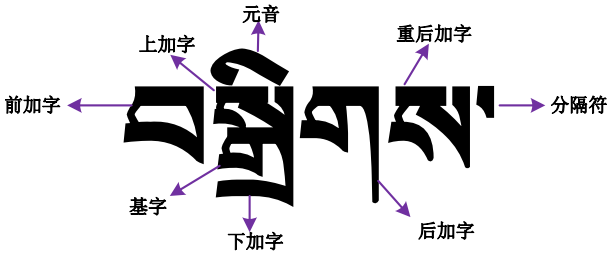



图 2. 藏文音节构成示例

的信息损失，从而影响最终的识别精度。此外，藏文句子中，部分功能性虚词 (多拉, 2011)如:“འི”、“འིས”、“ར”、“འོ”、“འམ”等通常不独立出现，而是附着在前面的藏文音节之后，形成了黏着现象。这种现象会改变原有藏文音节的结构，增加了TNER复杂度。例如，在句子“གླུ་བ་ནམ་མཁའི་གླུ་གཞས་ལ་ཉན་བཞིན་ལྷ་སར་འགྲོ།(意为“听着歌手南木夸的歌声去拉萨)”中，虚词“འི”黏着在人名实体“ནམ་མཁའ(南木夸)”之后形成黏着现象“ནམ་མཁའི”，而地名实体“ལྷ་ས(拉萨)”后面紧接的虚词“ར”形成黏着现象“ལྷ་སར”，都改变了原藏文音节的结构。另外，藏文人名如:“ཨོབྡན་བཀྲིས(俄见扎西)”是“ཨོ་རྒྱན་བཀྲ་ཤིས(俄见扎西)”的缩写，其中“ཨོ་རྒྱན་(俄见)”缩写为“ཨོབྡན(俄见)”，“བཀྲ་ཤིས(扎西)”缩写为“བཀྲིས(扎西)”，这样的缩写方式同样导致原藏文音节结构发生变化给TNER任务带来挑战。

针对目前藏文NER任务中藏文词汇信息和藏文音节部件信息被忽略的问题，本文提出一种有效结合藏文词汇信息和部件信息的藏文NER方法，采用基于Transformer编码器的平面网格结构，在藏文音节的基础上融合了藏文词汇信息。同时，再用一个Transformer编码器来提取藏文音节部件特征。通过交换两个编码器的查询值(Query)的方式实现了藏文多元数据特征的融合，从而有效地解决TNER任务中的挑战。主要贡献包括:

- 在藏文音节基础上融合了藏文词汇和藏文音节部件特征，采用多元藏文数据特征融合方法实现了藏文命名实体识别任务。
- 构建了藏文词汇词典和藏文音节部件词典，为后续相关研究提供了基础数据支持，利用交叉Transformer架构有效实现了对藏文多元数据的融合。
- 在数据集TibNER[1]上，本文所提出的方法相比基线模型表现出显著的提升，验证了该方法的有效性和实用性。

## 2 相关工作

近年来，命名实体识别领域取得了显著进展。 (Huang et al., 2015)首次将Bi-LSTM-CRF模型引入序列标注，取得了前所未有的成果。 (Zhang and Yang, 2018)提出了针对中文命名实体识别的Lattice-LSTM模型。该模型基于Bi-LSTM-CRF模型，通过综合编码中文词汇和字符信息，显著提高了对中文实体的识别准确性。 (Gui et al., 2019a)提出基于卷积神经网络(CNN)的中文字符级命名实体识别方法LR-CNN，采用重新思考机制整合词典，实现了更快速的识别速度和更好的性能。 (Gui et al., 2019b)提出了一种名为LGN的新型模型，采用图神经网络(GNN)作为基础，利用词典构建图结构，以捕获局部组合和全局语义。 (Yan et al., 2019)提出采用改进的Transformer编码器的NER模型TENER，证明了Transformer在NER任务中的有效性。 (Ma et al., 2019)提出的Soft_Lexicon方法简单有效地将词典融入字符表示中，使其易于迁移到任何中文命名实体识别模型上，实现了词汇词典的简化使用。 (Li et al., 2020)提出了基于Transformer的中文命名实体识别模型FLAT，将字符和词汇格子结构转换为跨度组成的平面结构，以实现对格子信息的充分利用，从而实现了出色的性能和效率。 (Wu et al., 2022)提出的NFLAT模型针对FLAT模型的冗余注意力计算进行了优化，显著降低了计算和内存成本，是用于中文命名实体识别的一种高效模型。 (Wu et al., 2021)提出了一种基于多元数据嵌入的交叉Transformer方法MECT，用于中文命名实体识别，通过融合汉字的结构信息来提高性能。此外，DAMO-NLP团队 (Wang et al., 2022)在MultiCoNER (Malmasi et al., 2022)共

[1]本文所用数据集已公开，地址：https://github.com/EJ-Cairang/TibNER_up


(c) 2024 中国中文信息学会计算语言学专业委员会

享任务中利用基于维基百科的多语言知识库，为命名实体识别模型提供相关的上下文信息，取得了10项胜利。接着，在此基础上，他们团队 (Tan et al., 2023)提出了U-RaNER检索增强系统，通过引入Wikidata知识库和注入方法来增强多语言细粒度NER任务的检索上下文，在MultiCoNER2 (Fetahu et al., 2023)共享任务的挑战中赢得了13项比赛中的9项冠军。

随着深度学习的发展，藏文命名实体识别(TNER)也取得了显著进展。 (朱亚军and 拥措, 2022)基于BiLSTM-CRF模型对藏文人名地名进行识别任务。 (环科尤et al., 2022)基于TS-BiLSTM-CRF网络识别格萨尔史诗文本中的六种格萨尔命名实体。 (洛桑嘎登et al., 2022)提出了一种融合藏文音节部件和藏文音节特征的模型SL-BiLSTM-CRF，验证了该方法对TNER任务的有效性。 (朱亚军et al., 2023)在基于音节的藏文BERT模型构建了BERT-BiLSTM-CRF模型，用于藏医药医学实体识别。 (徐泽辉et al., 2023)提出利用级联技术将任务分解为两个子任务的Cascade-BiLSTM-CRF模型，结合藏文预训练模型，以简化模型结构和缩短训练时间。 (格勒尼玛et al., 2023)将藏文音节特征和分词特征融合到了BiLSTM-CRF模型中，以预测藏文实体识别结果，验证了该方法的有效性。

当前用于TNER的基座模型主要采用LSTM网络结构，然而，LSTM网络在建立长距离依赖关系和并行处理能力方面存在一定限制。受到FLAT (Li et al., 2020)和MECT (Wu et al., 2021)方法在中文命名实体识别任务中的启发，本文提出了MECT-TL模型，采用基于Transformer的架构，在藏文音节基础上结合藏文词汇和音节部件信息，实现了藏文多元数据的融合。相较于现有TNER方法，本文方法具有更大的改进和创新。

## 3 MECT-TL:融合藏文词汇和音节部件特征的方法

### 3.1 整体框架方法

本文的总体模型框架**MECT-TL**(**MECT** for **T**ibetan **L**anguage)如图3所示，是对 (Wu et al., 2021)为中文命名实体识别所设计的MECT模型针对藏文特性进行改进，以适用于藏文命名实体识别任务。具体改进包括对藏文词汇信息融合和藏文音节部件信息融合。 MECT-TL模型

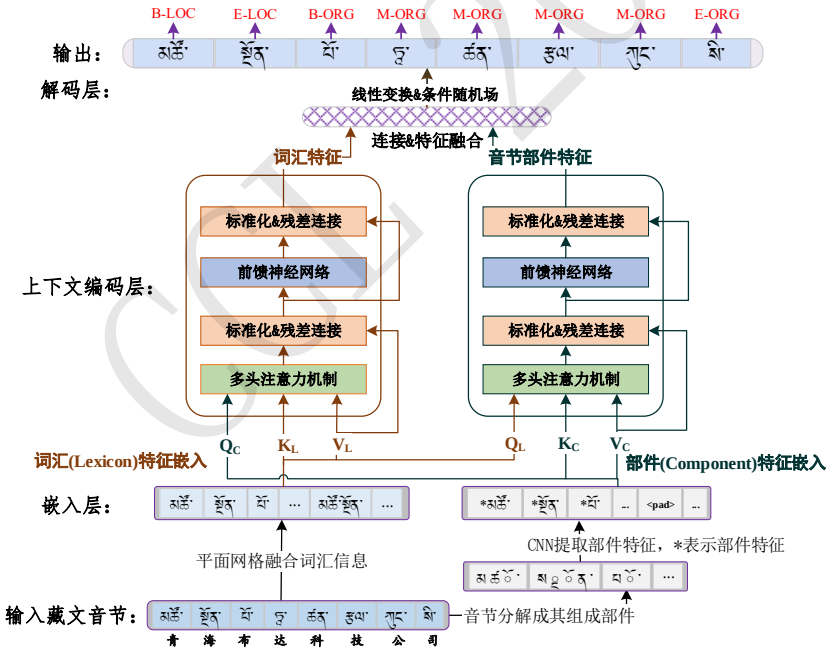



图 3. MECT-TL模型总体框架

包括两个Transformer编码器：融合词汇信息和提取音节部件信息。在这两个编码器中，输入文本序列分别经过词汇特征嵌入的Transformer编码器和音节部件特征嵌入的Transformer编码器，从而分别提取藏文词汇信息和音节部件信息。词汇特征嵌入的Transformer编码器在藏文音节的基础上融合相关的词汇信息，得到词汇特征；而音节部件特征嵌入的Transformer编码器则首先将藏文音节序列分解成对应的构成音节部件，随后通过CNN网络提取这些部件的特征，并将其送入专门用于整合藏文音节部件特征的编码器中，得到音节部件特征。通过交替整合词汇

信息和音节部件信息的Query以及连接操作，实现了对藏文音节、藏文词汇和音节部件的多元特征融合。最终，利用条件随机场(CRF)(?)进行解码，输出预测的标记序列，从而完成藏文命名实体识别任务。

## 3.2 词汇特征抽取

### 3.2.1 构建词汇词典

本文旨在探究融合不同粒度词汇对TNER的影响，本文采用Word2Vec处理了1.76GB的藏文文本，并构建了三种不同粒度的50维词向量词典，以捕捉词语之间的上下文关系：

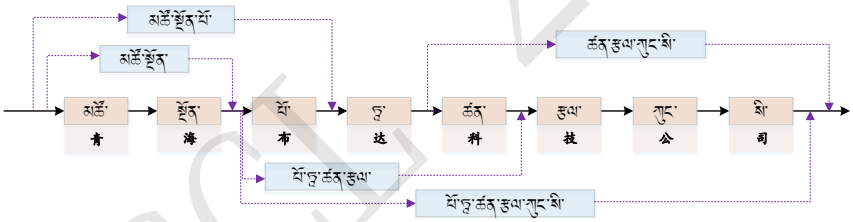
**音节词典**：包含22,973个藏文音节。有助于藏文的理解能力，改善语言模型。

**双音节词典**：由113,670个双音节藏文词汇构成。这增强了对双音节词汇语义的理解。

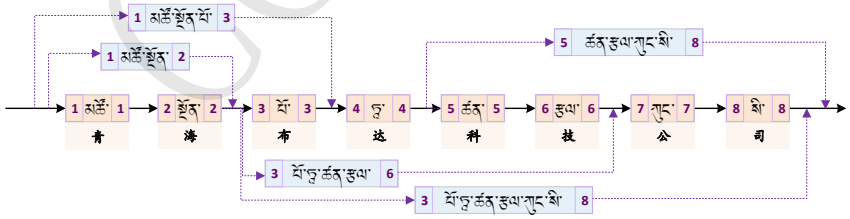
**综合词典**：包含单、双和多音节藏文词汇，共包含909,401个词汇。为复杂文本分析提供了强大支持。

### 3.2.2 融合词汇信息

基于藏文音节方法在捕捉词义和辨认词边界上的不足，缺乏关键信息导致产生不准确的实体边界预测。对输入藏文音节序列与藏文词典进行匹配，以便在词典中识别出所有潜在的词汇，从而识别出所有可能的词汇，并构建出词汇信息的晶体网格结构。如图4(a) 所示在这个结构中，每个节点不再是简单的藏文音节，而是扩展到了所有匹配到的潜在词汇。所形成的晶体结构是一个复杂的有向无环图，包括了整个藏文音节及其潜在的词汇。每个词汇在晶体中的位置是由它的首尾字符位置来明确界定的。为后续的实体辨识提供了清晰的边界指示。在这个多维结构中，识别特定词汇如人名、地名或机构名称至关重要，这是NER任务的核心。尽管晶体网格结构能够有效地融合藏文词汇信息，但其复杂性和动态性使得利用GPU并行计算能力变得困难，导致推理速度下降。图4(b)是带有位置索引的晶体结构。通过将晶体结构转化为平面结构图4(c)，平面结构中每个跨度对应于藏文音节或潜在藏文词汇及其在原始晶体网格中的位置，平面结构简化了表示方式。这种转换使模型能够更充分地利用网格信息，显著提高并行化能力，从而提升模型的性能和效率。




(a) 融合词汇信息的晶体网格结构




(b) 带有位置索引的晶体网格结构




(c) 融合词汇信息的平面网格结构

图 4. 融合词汇信息的晶体网络结构转平面网络结构

平面网格结构中，开始位置索引和结束位置索引表示每个跨度的第一个和最后一个藏文音节或藏文词汇的位置索引，它们表示文本单元在网格中的位置。对于藏文音节来说，它的开始位置编号和结束位置编号是一样的。有一种简单的算法可以将平面网格结构恢复到原始的晶体

网格结构。这样的处理既保持了数据的完整性，也便于充分利用Transformer的优势，利用位置编码将词汇信息在实体边界的预判中发挥作用，如图5所示。

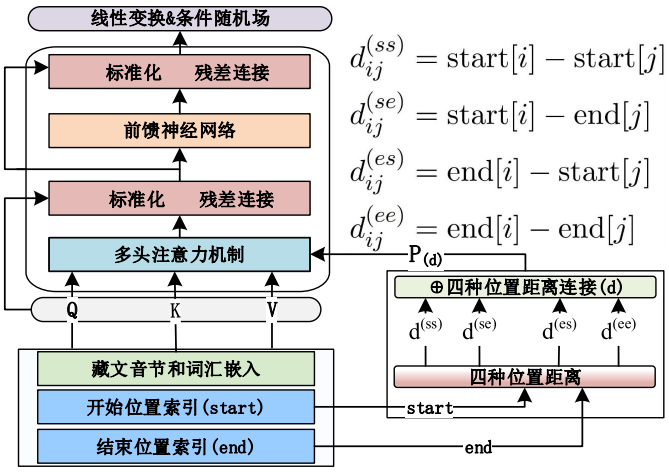



图 5. 平面网格结构的Transformer编码器

start[i]和end[i]表示跨度$X_i$的开始位置和结束位置。针对跨度$X_i$和$X_j$之间的关系，可以定义图中的$d_{ij}^{(ss)}$, $d_{ij}^{(se)}$, $d_{ij}^{(es)}$和$d_{ij}^{(ee)}$四种相对距离来描述它们之间的相互位置关系(相交、包含和相离)。$d_{ij}^{(ss)}$表示$X_i$和$X_j$的起始位置的距离，$d_{ij}^{(se)}$表示$X_i$的起始位置和$X_j$的结束位置的距离、$d_{ij}^{(es)}$表示$X_i$的结束位置和$X_j$的起始位置的距离、$d_{ij}^{(ee)}$表示$X_i$和$X_j$的结束位置的距离。相对位置编码的表达式为：

$$R_{ij} = \text{ReLU}\left(W_r\left(P_{d_{ij}^{(ss)}} \oplus P_{d_{ij}^{(se)}} \oplus P_{d_{ij}^{(es)}} \oplus P_{d_{ij}^{(ee)}}\right)\right) \tag{1}$$

其中$P_d$计算公式为：

$$P_d^{(2k)} = \sin\left(\frac{d}{10000^{2k/d_{\text{model}}}}\right), \quad P_d^{(2k+1)} = \cos\left(\frac{d}{10000^{2k/d_{\text{model}}}}\right) \tag{2}$$

这里的$P_d$指的是相对位置的嵌入表示。然后使用Transformer-XL (Dai et al., 2019)中的相对位置编码将$R_{ij}$加到Attention机制里面然后再使用自注意力的变形来计算跨度间相对位置编码的注意力分数，公式为：

$$A_{i,j}^* = W_q^T E_{x_i}^T E_{x_j} W_{k,E} + W_q^T E_{x_i}^T W_{k,R} R_{i-j} + u^T E_{x_j} W_{k,E} + v^T W_{k,R} R_{i-j} \tag{3}$$

最后将$A_{(i,j)}^*$代入到自注意力的计算公式中，可以得到整个序列的注意力值分布：

$$\text{Attention}(A, V) = \text{softmax}(A) \cdot V \tag{4}$$

$$A_{ij} = \left((Q_i + u)^T K_j\right) + \left((Q_i + v)^T R_{ij}^*\right) \tag{5}$$

$$[\text{Q}, \text{K}, \text{V}] = E_x[W_q, W_k, W_v] \tag{6}$$

其中$R_{ij}^* = R_{ij} \cdot W_R$，$u$和$v$是可学习的参数。这实现了在藏文音节基础上融合藏文词汇信息。

### 3.3 音节特征抽取

#### 3.3.1 音节部件词典构建

本文自行构建了一个包含18776个藏文音节及其构成部件对应的词典，几乎覆盖了藏文文本中可能出现的所有音节。每个音节都在基字(表1中粗体表示)的基础上被详细地分解为前加字、上加字、下加字、元音、后加字以及重后加字等组成部分。该词典精细地记录了每个音节的构成部件，为深入理解和分析藏文音节提供了有力的支持。表1展示了一些藏文音节及其对应构成部件的例子，这些示例反映了词典大致内容。表中符号“-”用以表示特定位置没有相应的音节部件。


(c) 2024 中国中文信息学会计算语言学专业委员会

| 藏文音节 | 前加字 | 上加字 | 基字 | 下加字 | 元音 | 后加字 | 重后加字 |
|---|---|---|---|---|---|---|---|
| བསྒྲིབས | བ | ས | ག | ྲ | ི | བ | ས |
| དཀྲོགས | ད | – | ཀ | ྲ | ོ | ག | ས |
| སྒྲེགས | – | ས | ག | ྲ | ེ | ག | ས |
| མཇུགས | མ | – | ཇ | – | ུ | ག | ས |
| འཁྱམ | འ | – | ཁ | ྱ | – | མ | – |
| རོབ | – | – | ར | – | ོ | བ | – |
| གསན | ག | – | ས | – | – | ན | – |
| བཀྲ | བ | – | ཀ | ྲ | – | – | – |
| ཞྭ | – | ལ | ཅ | – | ི | – | – |
| ཚ | – | – | ཚ | – | – | – | – |

表 1. 藏文音节及其对应构成部件示例

### 3.3.2 融合音节部件信息

藏文音节由1至7个音节部件构成，这些部件按照严格的藏文语法组合，赋予每个音节特定的语义。因此，音节部件序列在向量空间中能够有效地映射出音节的内在特征，对于理解音节的语义层面至关重要。为了更好地捕捉藏文音节序列的局部上下文特征，本文用一个基于卷积神经网络的模型。该模型能够通过局部特征提取强化音节嵌入的表征能力。如图6所示，展现了藏文音节部件特征提取的过程。图中的示例输入的藏文音节:“མཚོ(湖)”，可分解为三个部件：前加字:“མ”、基字:“ཚ”、上加字:“ོ”。这些部件随后通过卷积神经网络进行卷积运算、最大池化等操作，获得了带有特征的藏文音节部件特征。 然后将特征输送到Transformer编码器来获取藏

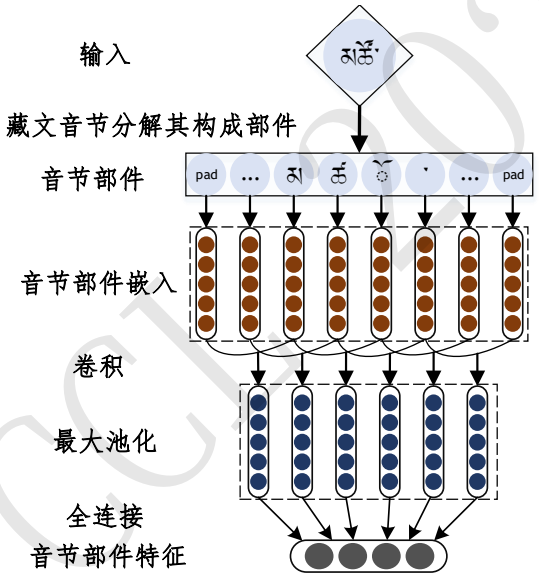



图 6. 基于卷积神经网络的音节部件特征抽取

文音节内部结构的补充语义信息。同时，利用了藏文音节的上下文和藏文词汇信息，丰富了藏文的语义表达，实现了多元特征的融合。输入$Q_L(Q_C)$、$K_L(K_C)$、$V_L(V_C)$是通过对藏文词汇融合和音节部件特征嵌入进行线性变换获得的：

$$\begin{bmatrix} Q_{L(C),i} \\ K_{L(C),i} \\ V_{L((C),i} \end{bmatrix}^T = E_{L(C),i} \begin{bmatrix} W_{L(C),Q} \\ I \\ W_{L(C),V} \end{bmatrix}^T \tag{7}$$

其中$E_L$和$E_C$分别是藏文词汇信息嵌入和音节部件信息嵌入，I是单位矩阵，每个W是一个可学习的参数。然后使用3.2.2中表述的相对位置编码来表示词的边界信息，并在模型中计算注意力得分：

$$\text{Attention\_}L\left(A_C, V_L\right) = \text{softmax}\left(A_C\right) V_L \tag{8}$$


(c) 2024 中国中文信息学会计算语言学专业委员会

$$\text{Attention\_}C\left(A_L, V_C\right) = \text{softmax}\left(A_L\right) V_C \tag{9}$$

$$A_{L(C),ij} = \left(Q_{L(C),i} + u_{L(C)}\right)^T K_{C(L),j} + \left(Q_{L(C),i} + v_{L(C)}\right)^T R^*_{L(C),ij} \tag{10}$$

其中u和v是方程(10)中注意力偏置的可学习参数，$A_L$ 是藏文词汇的注意力得分，$A_C$ 表示音节部件注意力得分。$R^*_{ij} = R_{ij} \cdot W_C$，其中$W_C$是可学习的参数。相对位置编码$R_{ij}$ 的计算如下：

$$R_{ij} = \text{ReLU}\left(W_r\left(p_{s_i - s_j} \oplus p_{e_i - e_j}\right)\right) \tag{11}$$

其中，s:start, e:end表示跨度的开始位置索引和结束位置索引。为了实现藏文词汇嵌入和音节部件特征嵌入所需的注意力偏置，该模型采用了随机注意力以提升性能。随机注意力能更好地适应两个子空间的分布。随机注意力是一个随机初始化的参数矩阵$B \in \mathbb{R}^{max_len \times max_len}$，它被添加到之前的注意力得分上，从而得到总的注意力得分：

$$\begin{aligned} V^*_L &= \text{Softmax}\left(A_C + B\right) V_L \\ V^*_C &= \text{Softmax}\left(A_L + B\right) V_C \end{aligned} \tag{12}$$

为了减少信息损失，直接连接藏文词汇特征和音节部件特征，并将它们输入到一个全连接层进行信息融合：

$$\text{Fusion}\left(V^*_L, V^*_C\right) = \left(V^*_C \oplus V^*_L\right) W^* + b \tag{13}$$

其中，符号$\oplus$的连接操作负责将来自音节部件和词汇信息的特征向量合并，形成一个全面的特征表示，这对于增强模型对藏文文本的解析能力至关重要。$W^*$和$b$是作为模型中的可学习参数，在训练过程中持续优化，以促进有效的特征融合。将融合后的特征被输入到条件随机场(CRF) (Lafferty et al., 2001)层进行识别任务。

## 4 实验与分析

### 4.1 实验数据与评估标准

#### 4.1.1 实验数据

本文利用藏文实体词典半自动地构建了藏文命名实体识别数据集——TibNER。为保证数据集的质量，我们对自动标注结果进行了人工校审。TibNER规模为55227句，标注的实体包括人名(PER)、地名(LOC)和组织机构名(ORG)，共计43678个实体。按照8:1:1的比例将TibNER数据集随机划分为训练集、验证集和测试集，数据详情如表2所示。

| 数据集 | 句子数量 | 实体类型及数量 | | | 实体总数 |
|---|---|---|---|---|---|
| | | 人名 | 地名 | 组织机构名 | |
| 训练集 | 44182 | 7790 | 15700 | 8764 | 32254 |
| 验证集 | 4997 | 933 | 1787 | 917 | 3637 |
| 测试集 | 6048 | 1068 | 2065 | 1155 | 4288 |
| **总计** | **55227** | **9791** | **19552** | **10836** | **40179** |

表 2. TibNER数据集统计信息

#### 4.1.2 评估标准

命名实体识别的评价标准对于评估模型性能至关重要。1996年 (Grishman and Sundheim, 1996)等提出了一套综合性评价标准，正确识别的实例需要系统同时正确识别其边界和类型。更具体地说，假阳性($F_P$)、假阴性($F_N$)和真阳性($T_P$)的数量用于计算精确度($P = \frac{T_P}{T_P+F_P} \times 100\%$)、召回率($R = \frac{T_P}{T_P+F_N} \times 100\%$)和F1值($F1 = \frac{2PR}{P+R} \times 100\%$)。

### 4.2 实验设置与对比系统

实验均在Linux 操作系统环境下进行，使用Tesla V100 GPU，CUDA 版本为12.2，PyTorch 版本为1.5.1。为了确定TibNER数据集上的最佳模型超参数配置，采用了SMAC算法来搜索最佳超参数。这种策略允许我们在预定义的参数空间内随机选择不同的超参数组合，随机搜索的超参数范围表3所示，并通过重复实验来评估它们对模型性能的影响。


(c) 2024 中国中文信息学会计算语言学专业委员会

| 超参数 | 范围 |
|---|---|
| output_dropout | [0.1, 0.2, 0.3] |
| lattice_dropout | [0.1, 0.2, 0.3] |
| radical_dropout | [0.1, 0.2, 0.3, 0.4] |
| warm_up | [0.1, 0.2, 0.3] |
| $d_{head}$ | [16, 20] |
| $d_{model}$ | [128, 160] |
| lr | [1e-3, 25e-4] |
| radical_lr | [6e-4, 25e-4] |
| momentum | [0.85, 0.97] |

表 3: 超参数范围

MECT-TL与主流的TNER模型CRF、BiLSTM、BiLSTM、TENER进行了评估:
**CRF:**考虑了标签序列之间的依赖关系，通过最大化条件概率来确定最佳标签序列。
**BiLSTM:**通过双向LSTM网络捕捉序列数据中的长依赖关系和上下文信息。
**BiLSTM-CRF:**结合了双向长短期记忆网络(BiLSTM)和条件随机场(CRF)的模型，既能捕捉序列数据中的长依赖关系和上下文信息，又考虑了标签序列之间的依赖关系，通过最大化条件概率来确定最佳标签序列。目前TNER任务中常用的基准模型。
**TENER:**专门为命名实体识别任务开发的中文NER模型，在Transformer编码器中巧妙融合位置信息的自注意力机制。这种机制有效地捕捉了文本序列中的长距离上下文依赖关系，使得TENER能够识别和处理文本中的复杂地长依赖关系。

#### 4.2.1 主实验

在TibNER数据集上与主流TNER模型进行了对比实验，结果见表4所示。

| TNER模型 | P(%) | R(%) | F1(%) |
|---|---|---|---|
| CRF | 81.49 | 84.13 | 82.79 |
| BiLSTM | 74.02 | 83.99 | 78.69 |
| BiLSTM-CRF | 83.84 | 84.47 | 84.15 |
| TENER | 84.96 | 85.25 | 85.11 |
| **MECT-TL(Our)** | **88.49** | **90.11** | **89.29** |

表 4. TNER模型性能评估实验

本文模型MECT-TL在F1值上达到了89.29%，表现最佳。相比当前主流的TNER基准模型BiLSTM-CRF，本文模型提高了5.14% 的F1值；而相较于基于Transformer架构的TENER模型，则提高了4.18% 的F1值。MECT-TL模型在TNER任务中表现出卓越的性能。该模型在藏文音节的基础上融合了藏文词汇和音节部件信息，为其提供了丰富的多元数据特征，使其能够全面获取知识。可见，藏文词汇增强和音节部件增强方法对提升TNER任务性能具有显著的增益效果。

#### 4.2.2 消融实验

通过逐步融合藏文的单音节词典、双音节词典以及包含单音节、双音节及更多音节的综合词典，以评估不同粒度的词汇信息对模型性能的影响。接着，比较仅融合音节部件的效果与在最佳增益效果的综合词典上再融合音节部件的模型性能。实验结果表5所示。表中“✓”表示有对应的词典融合，“✗”表示没有对应的词典融合。

实验结果表明，融合不同粒度的词汇信息对提升模型的藏文命名实体识别性能具有显著影响。其中，融合双音节词典比融合单音节词典具有更大的增益效果，这表明词汇有助于正确判定实体边界，并捕捉更复杂的语义信息。综合词汇信息融合后，性能达到最优，反映了融合多音节及更丰富词汇信息对模型理解和处理藏文文本的重要性。特别是对于藏文这种结构复杂的语言，有效的词汇信息融合显著提升了命名实体识别的准确度。再通过比较仅融合音节部件的效果与在最佳增益效果的综合词典上再融合音节部件的模型实验，可以发现融合词汇的效果比


(c) 2024 中国中文信息学会计算语言学专业委员会

| 藏文词汇词典 | | | 音节部件词典 | P(%) | R(%) | F1(%) |
|---|---|---|---|---|---|---|
| 单音节 | 双音节 | 综合 | | | | |
| ✓ | ✗ | ✗ | ✗ | 84.14 | 85.61 | 84.67 |
| ✗ | ✓ | ✗ | ✗ | 85.06 | 86.45 | 85.75 |
| ✗ | ✗ | ✓ | ✗ | 87.15 | 89.98 | 88.54 |
| ✗ | ✗ | ✗ | ✓ | 84.83 | 86.04 | 85.43 |
| | | ✓ | ✓ | **88.49** | **90.11** | **89.29** |

表 5. 消融实验结果

融合音节部件的增益效果更为显著。同时融合两者将能取得最佳的结果，说明融合多元特征的方法对TNER的有效性，这为未来藏文处理模型的优化和其他低资源语言的命名实体任务提供了启示。

### 4.2.3 推理效率对比实验

为了验证MECT-TL的计算效率，在TibNER数据集上进行了推理速度的对比实验，包括非批量处理(batch=1)和批量处理(batch=16)条件。结果如图7所示，其中我们将BiLSTM-CRF设为基准速度(值为1)以便于对比。

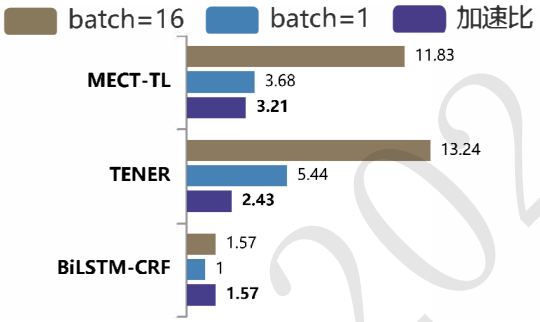



图 7. 不同模型计算效率对比实验

在非批量条件下，本文模型MECT-TL尽管在藏文音节特征基础上多融合了两种不同特征数据，但相较于BiLSTM-CRF的基准模型，其推理速度快了3.68倍，相比TENER模型慢了1.76倍。而在批量条件下，BiLSTM- CRF的加速比约为1.75，TENER的加速比约为3.06，表明模型在较大批处理下有一定的性能提升。而本文模型MECT-TL加速比达到了3.76，表现出了最佳的模型效率。虽然TENER推理速度比MECT-TL快，但MECT-TL模型在较大批处理下展现出了更为明显的性能提升。这表明本文模型在推理速度方面展现了高效的并行处理潜力，并未因多种特征融合而导致显著的推理速度下降。进一步突显了其在结构简洁性和计算效率方面的设计成功。MECT-TL模型表现出了强大的数据并行处理能力和快速有效的推理长距离依赖能力。

## 5 总结

本文针对传统藏文命名实体识别(TNER)方法的只依赖藏文音节层面嵌入的问题，通过采用交叉Transformer的先进方法，在藏文音节上融合了藏文词汇信息和藏文音节部件信息，实现了多元藏文数据特征融合表示。藏文词汇的融合使得模型更深入理解藏文词汇的语义信息。模型通过计算字词之间包含、相交、相离的距离关系得到的词汇注意力分数在判断文本中的实体边界起到关键的辅助作用，提高了实体识别的准确率。藏文音节部件信息的融合让模型利用这些承载语义信息的部件，提取到了藏文音节的内部结构，从而提高了TNER识别任务的性能。通过一系列细致的实验和分析，展示了所提方法在提升TNER任务的准确率和效率方面的优势。

## 致谢

感谢所有匿名审稿人的宝贵意见。本项研究成果受国家社会科学基金重大项目


(c) 2024 中国中文信息学会计算语言学专业委员会

（22&ZD035）、中国语言资源保护工程（课题编号：YB2404A003）、中央民族大学项目(GRSCP202316, 2023QNYL22, 2024GJYY43)资助。

## 参考文献

Bogdan Babych and Anthony Hartley. 2003. Improving machine translation quality with automatic named entity recognition. In *Proceedings of the 7th International EAMT workshop on MT and other language technology tools, Improving MT through other language technology tools, Resource and tools for building MT at EACL 2003*.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

Dennis Diefenbach, Vanessa Lopez, Kamal Singh, and Pierre Maret. 2018. Core techniques of question answering systems over knowledge bases: a survey. *Knowledge and Information systems*, 55:529–569.

Besnik Fetahu, Sudipta Kar, Zhiyu Chen, Oleg Rokhlenko, and Shervin Malmasi. 2023. Semeval-2023 task 2: Fine-grained multilingual named entity recognition (multiconer 2). *arXiv preprint arXiv:2305.06586*.

Ralph Grishman and Beth M Sundheim. 1996. Message understanding conference-6: A brief history. In *COLING 1996 volume 1: The 16th international conference on computational linguistics*.

Tao Gui, Ruotian Ma, Qi Zhang, Lujun Zhao, Yu-Gang Jiang, and Xuanjing Huang. 2019a. Cnn-based chinese ner with lexicon rethinking. In *ijcai*, volume 2019.

Tao Gui, Yicheng Zou, Qi Zhang, Minlong Peng, Jinlan Fu, Zhongyu Wei, and Xuan-Jing Huang. 2019b. A lexicon-based graph neural network for chinese ner. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 1040–1050.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Deniz Karatay and Pinar Karagoz. 2015. User interest modeling in twitter with named entity recognition. In *5th Workshop on Making Sense of Microposts*.

Mahboob Alam Khalid, Valentin Jijkoun, and Maarten De Rijke. 2008. The impact of named entity normalization on information retrieval for question answering. In *European Conference on Information Retrieval*, pages 705–710. Springer.

John Lafferty, Andrew McCallum, Fernando Pereira, et al. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Icml*, volume 1, page 3. Williamstown, MA.

Xiaonan Li, Hang Yan, Xipeng Qiu, and Xuanjing Huang. 2020. Flat: Chinese ner using flat-lattice transformer. *arXiv preprint arXiv:2004.11795*.

Ruotian Ma, Minlong Peng, Qi Zhang, and Xuanjing Huang. 2019. Simplify the usage of lexicon in chinese ner. *arXiv preprint arXiv:1908.05969*.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022. Multiconer: A large-scale multilingual dataset for complex named entity recognition. *arXiv preprint arXiv:2208.14536*.

Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M Marlin. 2013. Relation extraction with matrix factorization and universal schemas. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 74–84.

Burr Settles. 2004. Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications (NLPBA/BioNLP)*, pages 107–110.


(c) 2024 中国中文信息学会计算语言学专业委员会

Zeqi Tan, Shen Huang, Zixia Jia, Jiong Cai, Yinghui Li, Weiming Lu, Yueting Zhuang, Kewei Tu, Pengjun Xie, Fei Huang, et al. 2023. Damo-nlp at semeval-2023 task 2: A unified retrieval-augmented system for multilingual named entity recognition. *arXiv preprint arXiv:2305.03688*.

Zhibo Wang, Xiaohui Cui, Lu Gao, Qi Yin, Lei Ke, and Shurong Zhang. 2016. A hybrid model of sentimental entity recognition on mobile social media. *EURASIP Journal on Wireless Communications and Networking*, 2016:1–12.

Xinyu Wang, Yongliang Shen, Jiong Cai, Tao Wang, Xiaobin Wang, Pengjun Xie, Fei Huang, Weiming Lu, Yueting Zhuang, Kewei Tu, et al. 2022. Damo-nlp at semeval-2022 task 11: A knowledge-based system for multilingual named entity recognition. *arXiv preprint arXiv:2203.00545*.

Shuang Wu, Xiaoning Song, and Zhenhua Feng. 2021. Mect: Multi-metadata embedding based cross-transformer for chinese named entity recognition. *arXiv preprint arXiv:2107.05418*.

Shuang Wu, Xiaoning Song, Zhenhua Feng, and Xiao-Jun Wu. 2022. Nflat: Non-flat-lattice transformer for chinese named entity recognition. *arXiv preprint arXiv:2205.05832*.

Hang Yan, Bocao Deng, Xiaonan Li, and Xipeng Qiu. 2019. Tener: adapting transformer encoder for named entity recognition. *arXiv preprint arXiv:1911.04474*.

Yue Zhang and Jie Yang. 2018. Chinese ner using lattice lstm. *arXiv preprint arXiv:1805.02023*.

多拉. 2011. 藏语语义理解中功能性虚词研究. 西藏大学学报*(*社会科学版*)*, 26:106–112.

徐泽辉, 珠杰, 许泽洲, 汪超, 严松思, and 刘亚姗. 2023. 结合级联技术的藏文预训练命名实体识别模型. 中文信息学报, 37:23–28.

朱亚军and 拥措. 2022. 基于深度学习的藏文人名地名识别. 信息与电脑*(*理论版*)*, 34:66–68.

朱亚军, 拥措, and 尼玛扎西. 2023. 基于藏文bert的藏医药医学实体识别. 计算机与现代化, pages 43–48.

格勒尼玛, 群诺, 项秀才让, 洛桑嘎登, and 尼玛扎西. 2023. 结合分词特征的藏文命名实体识别方法. 高原科学研究, 7:106–114.

洛桑嘎登, 群诺, 索南尖措, and 仁增多杰. 2022. 融合音节部件特征的藏文命名实体识别方法. 厦门大学学报*(*自然科学版*)*, 61:624–629.

环科尤, 华却才让, 才让当知, and 多杰才让. 2022. 基于深度学习的格萨尔史诗命名实体识别研究. 中文信息学报, 36:46–53.

CCL 2024

# PGA-SciRE：基于大语言模型的数据增强框架进行科学领域的关系抽取

周洋[1]，单世民[1*]，魏宏夔[2]，赵哲焕[1*]，冯文铄[1]
[1]大连理工大学 软件学院，大连市 116620
[2]北京电子工程总体研究所 复杂产品智能制造系统技术国家重点实验室，北京市 100854
{zy22217034, fengws}@mail.dlut.edu.cn
weihongkui89@hotmail.com
{ssm, z.zhao}@dlut.edu.cn

## 摘要

关系提取旨在识别文本中提到的实体对之间的关系。大语言模型的进步对自然语言处理任务产生了巨大的影响。在这项工作中，我们针对科学领域的关系抽取任务，提出一个名为PGA的数据增强框架，用于提升模型在科学领域的关系抽取的性能。框架引入了两种数据增强的方式，利用大语言模型通过转述原训练集样本，得到句意相同但具备不同表述和形式的伪样本。以及指导大语言模型根据原训练集样本的关系和实体标签，生成暗含对应标签信息的句子。这两种伪样本分别与原数据集共同参与关系抽取模型的训练。实验中PGA框架提高了三个主流模型的科学领域内关系抽取的F1分数。同时，使用大语言模型获得样本也能有效减少人工标注数据的成本。

**关键词：** 数据增强；关系抽取

# PGA-SciRE: Harnessing LLM on Data Augmentation for Enhancing Scientific Relation Extraction

**Yang Zhou[1], Shimin Shan[1*], Hongkui Wei[2], Zhehuan Zhao[1*], Wenshuo Feng[1]**
[1]Software school, Dalian University of Technology, Dalian 116620
[2]State Key Laboratory of Intelligent Manufacturing System Technology,
Beijing Institute of Electronic System Engineering, Beijing 100854
{zy22217034,fengws}@mail.dlut.edu.cn
weihongkui89@hotmail.com
{ssm, z.zhao}@dlut.edu.cn

## Abstract

Relation Extraction aims at recognizing the relation between pairs of entities mentioned in a text. Advances in LLMs have had a tremendous impact on NLP tasks. In this work, we propose a textual data augmentation framework called PGA for improving the performance of models for relation extraction task in the scientific domain. The framework introduces two ways of data augmentation, utilizing a LLM to obtain pseudo-samples with the same sentence meaning but with different representations and forms by paraphrasing the original training set samples. As well as instructing LLM to generate sentences that implicitly contain information about the corresponding labels based on the relation and entity of the original training set samples. These two kinds of pseudo-samples participate in the training of the relation extraction model together with the original dataset, respectively. The PGA framework in the experiment improves the F1 scores of the three mainstream models for relation extraction within the scientific domain. Also, using a large language model to obtain samples can effectively reduce the cost of manually labeling data.

*共同通讯作者 Co-corresponding Author.
©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

**Keywords:** Data Augmentation , Relation Extraction

## 1 引言

大语言模型(Large Language Model, LLM)的跨越式进展极大地提高了各项自然语言处理(Natural Language Processing, NLP)任务的性能和效率，而新的上下文学习方法的提出也为利用大语言模型解决各种NLP任务创造了新的机遇(Brown et al. (2020))。GPT系列是目前最流行的LLM，以其准确理解用户意图和产生类似人类反馈的能力而闻名。大量工作研究了LLM的性能、推理和解释能力，提出了系统的分析(Li et al. (2023a))。Wei et al. (2023)通过研究LLM在各种信息抽取(Information Extraction, IE)任务上的性能，对LLM的能力进行全面研究和详细分析。虽然LLM的性能无法与当前的SOTA模型相提并论，但证明了LLM拥有与众不同的解决NLP各种任务的能力。Han et al. (2023)提出新的对LLM的IE评价标准，即一个软匹配的评估策略，以更准确地反映LLM的性能。这些工作都证明了LLM在NLP任务中的巨大潜能。

关系抽取(Relation Extraction, RE)是从非结构化文本中获取结构化知识，是识别文本中实体之间的语义关系的任务。通常基于跨度的关系抽取方法是学习实体的跨度，再对枚举出的跨度进行多分类(Eberts and Ulges (2019))，亦或是用精心设计的特定的标记对将带有标签的实体包裹起来，以便预训练模型更好地学习实体之间的关系(Ye et al. (2022); Zhong and Chen (2021))。

一直以来，为NLP各种任务获取大规模且高质量的带有标注信息的数据是一件非常有挑战性、代价极高的事情(Beltagy et al. (2019))。特别是在科学知识图谱领域中，因为相对通用领域的的数据集，科学文献数据集往往会面临着独特的挑战。比如科学数据集可能会有更多的噪音；包含大量专业词汇、以及特殊概念的特有缩写，使得模型难以学习和泛化；由于整理它们需要花费大量时间和人力成本，因此这些数据集往往缺乏大量标注。

在过去的工作中，基于LLM来解决关系抽取任务的方法一般分为两个方向。第一种是利用上下文学习(In-Context Learning, ICL)和思维链(Chain of Thought, CoT)的方法，通过将关系抽取视为条件文本生成任务(Wan et al. (2023); Ma et al. (2023a))或问答任务(Zhang et al. (2023); Li et al. (2023b))。输入测试样例、示例、任务描述，由LLM直接给出答案。这些方法主要应用在Low-shot设置下，同时由于LLM输出的语言的灵活性，导致评估其输出的结果也成为了一种挑战(Wadhwa et al. (2023))。同时，由于科学领域的专业概念每年被持续地定义和提出，而训练LLM的语料库中并不一定包含这些缩写和概念，所以这些方法并不一定擅长工作于科学领域的关系抽取。

另外一种方法则是从模型的训练数据入手。除开模型开发，关系抽取始终涉及一定程度的标注数据。不仅是few-shot的设置下模型，全样本训练设置下的模型的性能也很大程度上被训练数据的质量和数量所影响，但是标注科学领域的数据所需的成本却又较高，这是一个在实践中非常常见的问题。LLM为解决这一问题带来新的方法，为生成标注数据提供了可能(Møller et al. (2024))。过去基于LLM的数据增强方法主要聚焦于解决文本分类(Dai et al. (2023); Piedboeuf and Langlais (2023); Tang et al. (2023))，与之不同的是，关系抽取通常包含更加丰富的预定义标签信息(Xu et al. (2023c))，以及更广的分类空间，并且科学领域数据集上的关系抽取则是更具有挑战性。此外，这些工作都致力于few-shot设置下的RE，数据增强只需要产生少量的样本就能让模型的性能得到提升。与之相反的是，通常的全样本模型在实践中需要的数据的规模和质量要求远远超过few-shot设置。

为此，我们从数据方面入手，将LLM难以理解的科学领域的专有词汇作为LLM的输入信息，弥补其在这一领域的知识，使其合成额外的域内数据样本。这项工作提出了名为PGA(Paraphrasing and Generating Augmentation)用于加强科学领域的关系抽取的框架，框架利用大语言模型GPT-3.5通过转述和生成两种方法来合成伪数据以加强关系抽取模型的性能，其贡献如下。

1. 这项工作的研究表明，LLM可以通过我们设计的简单而有效的提示，能够在无需人工标注的情况下生成较优的、带有标注的科学领域的关系抽取伪样本。

2. 本文提出的PGA框架通过转述和生成的方式获得的伪样本分别能够让几个主流关系抽取模型的F1分数得到不同程度的提高。


(c) 2024 中国中文信息学会计算语言学专业委员会

## 2 相关工作

**关系抽取** 关系提取旨在识别文本中提到的实体对之间的关系。主流的方法可分为流水线法和联合法。(1)流水线方法依次处理两个子任务，命名实体识别(Named Entity Recognition, NER)和关系提取，但会存在错误传播的问题(Ye et al. (2022); Zhong and Chen (2021);Miwa and Bansal (2016))。(2)联合方法可以通过同时处理两个子任务来缓解此问题。Yan et al. (2021); Wang et al. (2021); Wang and Lu (2020)将两个任务都作为表项的标记来处理。Sun et al. (2019); Fu et al. (2019); Nguyen et al. (2021)则利用图卷积网络(Graph Convolution Networks, GCN)从实例依赖图入手，设计模型进行联合关系抽取。Yan et al. (2023)利用超图神经网络(Hypergraph Graph Neural Networks, HGNN)来进行高阶建模，并以Ye et al. (2022)的流水线模型为基础，设计联合的关系抽取模型。Kong and Xia (2023)提出了联合抽取的协同关注网络，共同关注模块捕捉两个子任务之间的双向互动，利用实体信息进行关系预测。Chen et al. (2022); Zhang et al. (2022)使用位置感知注意力设计联合提取模型。Zaratiana et al. (2024)基于文本跨度的生成线性化图，在跨度和关系类型的动态词汇表上采用了带有指向机制的转换器编码器-解码器架构。此外，Ren et al. (2023)使用协方差优化方法将特征的协方差最小化，增强了各流水线和联合模型的特征的表征能力。

在科学领域关系抽取方面，Luan et al. (2018)提出了一个具有共享跨度表示的框架，减少了任务之间的级联错误。Jiang et al. (2020)利用句法规则作为远距离监督的一种形式来连接相关的科学术语对，训练分类器来进一步识别每对术语对的关系类型。Santosh et al. (2021)用词性标记(POS tags)的嵌入对transformer生成标记的上下文嵌入进行丰富，改进了科学关系抽取。

**LLM改善NLP任务** 最近，LLM在信息抽取任务的各方面指标上都展现出了可观的性能，这也为NLP任务的进步注入了新的活力(Li et al. (2023a); Zhu et al. (2024))。过去的工作除了研究ChatGPT的信息抽取性能之外，还对LLM独有的评价标准进行了研究，以求更准确地反映其性能(Han et al. (2023); Wadhwa et al. (2023))。此外，Wei et al. (2023)为关系抽取任务提出一个两阶段的ChatGPT框架，性能与过去的微调模型而言较有竞争力。Yuan et al. (2023)研究了ChatGPT在zero-shot时序关系提取方面的能力，发现其性能与有监督的方法有很大的差距。Gao et al. (2023b)同样证明了ChatGPT用于事件提取还有一定挑战。在医学领域方面，Agrawal et al. (2022)研究的临床提取任务中，GPT-3的表现超过了现有的zero-shot和few-shot的基线。Jimenez Gutierrez et al. (2022)研究了相较于微调较小的PLM，GPT-3 ICL方法在两个生物医学信息提取任务上性能明显不足。

在关系抽取方面，Ma et al. (2023a)提出了一种使用LLM进行zero-shot关系提取方法，利用特定任务和知识诱导LLM生成证据，并将其纳入CoT提示以进行关系提取。Wan et al. (2023)通过在演示检索中加入任务感知表征，以及使用金标签诱导推理逻辑来丰富演示，从而进行few-shot关系抽取。Zhang et al. (2023)使用LLM将关系抽取与问题解答(QA)结合起来，以解决zero-shot关系抽取。Li et al. (2023b)利用LLM递归的输入提示，将zero-shot关系抽取输入转换为有效的QA格式。Tao et al. (2024)利用LLM和ICL通过基于示例的详细推理来显著增强few-shot关系提取，图形推理方法将关系提取分解为连续的子任务，提高了处理复杂关系数据的精度和适应性。Li et al. (2024)将LLM与检索语料协同作用，从而实现相关检索和可靠的上下文推理。

另外，模型协同(Model Collaboration)也能提高特定任务的性能。综合LLM和基于微调的传统模型共同解决关系抽取。Ma et al. (2023b)提出了一种自适应“过滤-重排”(Filter-then-Rerank)范式，LLM对小语言模型(Small Language Model, SLM)识别出的一小部分困难样本进行重排，以提升few-shot关系抽取性能。Xu et al. (2023b)提出了超级上下文学习(SuperICL)，允许LLM与局部微调的SML一起工作，提高性能，同时解决ICL的不稳定问题。Tang et al. (2024)采用“训练-指导-预测”(Training-Guide-Predict)策略，预训练语言模型(Pre-trained Language Model, PLM)充当导师，将任务知识传输给LLM，并指导LLM进行关系抽取。

**数据增强** RE任务下的数据增强是在不改变原始数据标签的设置下，增加数据集的规模和多样性，帮助模型更好地学习实体以及关系标签的表达方式。合成新的优质数据能改善和加强RE模型的训练，提高模型的泛化能力。常见的文本数据增强方法包括同义词替换(Hsu et al. (2021); Wei and Zou (2019))、随机插入、随机交换、随机删除(Belinkov and Bisk (2018))和回译(Sennrich et al. (2016); Hayashi et al. (2018))等。这些方法对原始文本进行变换，得到新的

第二十三届中国计算语言学大会论文集，第352页-第369页，太原，中国，2024年7月25日至28日。
卷1：主会论文
(c) 2024 中国中文信息学会计算语言学专业委员会

句子。

转述式数据增强(Paraphrasing-based Data Augmentation)方法是对原始文本进行重写或重新表述，以生成具有相同意义但表达形式不同的新文本，转述技术被广泛应用于各NLP任务(Kumar et al. (2019); Okur et al. (2022))，包括关系抽取(Yu et al. (2020))。由于LLM的兴起，转述技术也因此得到蓬勃发展，Piedboeuf and Langlais (2023)研究了ChatGPT在文本分类任务上的转述数据的性能。

生成式数据增强(Generating-based Data Augmentation)方法要求生成的数据需要尽可能接近原数据集的数据分布，并且保持与原始数据相似的信息。Meng et al. (2022); Gao et al. (2023a)聚焦于利用预训练语言模型(BERT)或生成式预训练语言模型(GPT-2)进行生成数据，但受数据多样化和需不能完全脱离原数据集分布的双重限制，它们的模型设计和微调往往具有挑战性(Pouran Ben Veyseh et al. (2023))。然而，基于GPT3.5系列的方法不再需要微调，更接近于一种端到端的方式进行生成数据。特别是在低资源领域(Xu et al. (2023a))，GPT3.5经过海量语料库训练，拥有广阔的数据储备，这些先验知识极大促进了生成数据的准确性，以提高文本分类任务(Dai et al. (2023); Piedboeuf and Langlais (2023); Tang et al. (2023))和关系抽取任务的性能(Yoo et al. (2021); Xu et al. (2023c))。

## 3 方法

### 3.1 任务定义

**关系抽取(Relation Extraction, RE)** 关系抽取任务的目标是根据给定的句子，抽取句子中（如果存在）提到的若干成对实体之间的关系。通常关系抽取数据集由句子语料库、实体对和关系类型构成。一个关系抽取样本示例中包含一个由多个单词$w$组成的句子$S = \{w_l\}_{l=1}^{L}$，以及$S$所提及的成对的头实体$e_s$和尾实体$e_o$，其中$e_s, e_o \in \mathbb{E}$和关系$r \in \mathbb{R}$，其中$\mathbb{E}$和$\mathbb{R}$分别代表一组预定义的实体类型和关系类型。具体来说，关系抽取任务是指给定一个句子$S$，模型需要从$\mathbb{R}$中，预测出$S$中包含的$e_s$ 和$e_o$之间的对应关系（如果存在），得到关系抽取三元组$(e_s, r, e_o)$。即模型需要学习每个训练样本$\mathcal{D} = \{(S, e_{s_i}, e_{o_i})\}_{i=1}^{N}$，来预测其对应的关系标签$r_i$：

$$p_{LM}\left(r|S, e_s, e_o\right) = \prod_{i=1}^{N} p\left(r_i|S, e_{s_i}, e_{o_i}\right) \tag{1}$$

模型预测可能的关系集合的概率分布，并根据每个训练样本$\mathcal{D}$来预测概率最高的关系：

$$r^* = \arg\max_{r \in \mathbb{R}} \prod_{i=1}^{N} p\left(r_i|S, e_{s_i}, e_{o_i}\right) \tag{2}$$

**跨度级的实体和关系抽取(Entity and Relation Extraction, ERE)** 本工作旨在解决基于跨度的实体和关系抽取任务，即上述RE任务中给定的所有实体信息由模型中的NER模块预测获得。更具体地说，基于跨度指的是形式上，给定的句子$S$由多个分词(token)构成：$S = \{t_i\}_{i=1}^{n}$。跨度定义为连续的token序列，跨度有开始和结束的索引代表跨度的边界，定义$Span = \{x_1, x_2, x_3, ..., x_m\}$为所有可能的跨度，实体定义为标有实体类型的跨度，关系定义为标有关系类型的实体跨度对。对NER任务来说，模型需要预测每个跨度的实体类型，即映射每个跨度到实体类型中：$y_e(x_i) \in \mathbb{E}$代表跨度是实体，并具有相应实体类型，而$y_e(x_i) \notin \mathbb{E}$代表跨度不是实体，即模型NER部分输出为$Y_e = \{(x_i, e) : x_i \in Span, e \in \mathbb{E}\}$。对RE任务来说，模型需要对每一对实体跨度$x_s, x_o \in Span$进行预测其关系类型，即映射每对实体跨度到关系类型中：$y_r(x_s, x_o) \in \mathbb{R}$代表代表这对实体跨度存在关系，并具有相应关系类型，$y_r(x_s, x_o) \notin \mathbb{R}$代表它们没有关系，即模型RE部分输出为$Y_r = \{(x_s, x_o, r) : s_s, s_o \in Span, r \in \mathbb{R}\}$。

### 3.2 方法概述

如图1所示，这项工作提出了一个科学领域关系抽取的数据增强框架PGA(Paraphrasing-based and Generating-based Augmentation)，通过生成和转述的方式来指导LLM生成伪样本来提升关系抽取模型的性能。更具体地来说，首先针对LLM易理解的样本数据格式，分别设计用于转述和生成方式的数据增强地提示语，提示由基础的3部分构成：Task Formulation、Demonstration Instruct、Sample Input。需要注意的是，生成式数据增强方法还额外包


(c) 2024 中国中文信息学会计算语言学专业委员会

含Label Interpretation。通过这些精心设计的提示，指导LLM生成大量带有标签信息的伪样本。接着，经过数据后处理步骤，过滤错误样本，以控制样本质量，并将伪样本变换成原有的数据格式，利用这些额外的伪样本与原有的训练数据集或验证集共同对各骨干RE模型进行微调，让模型提升性能。

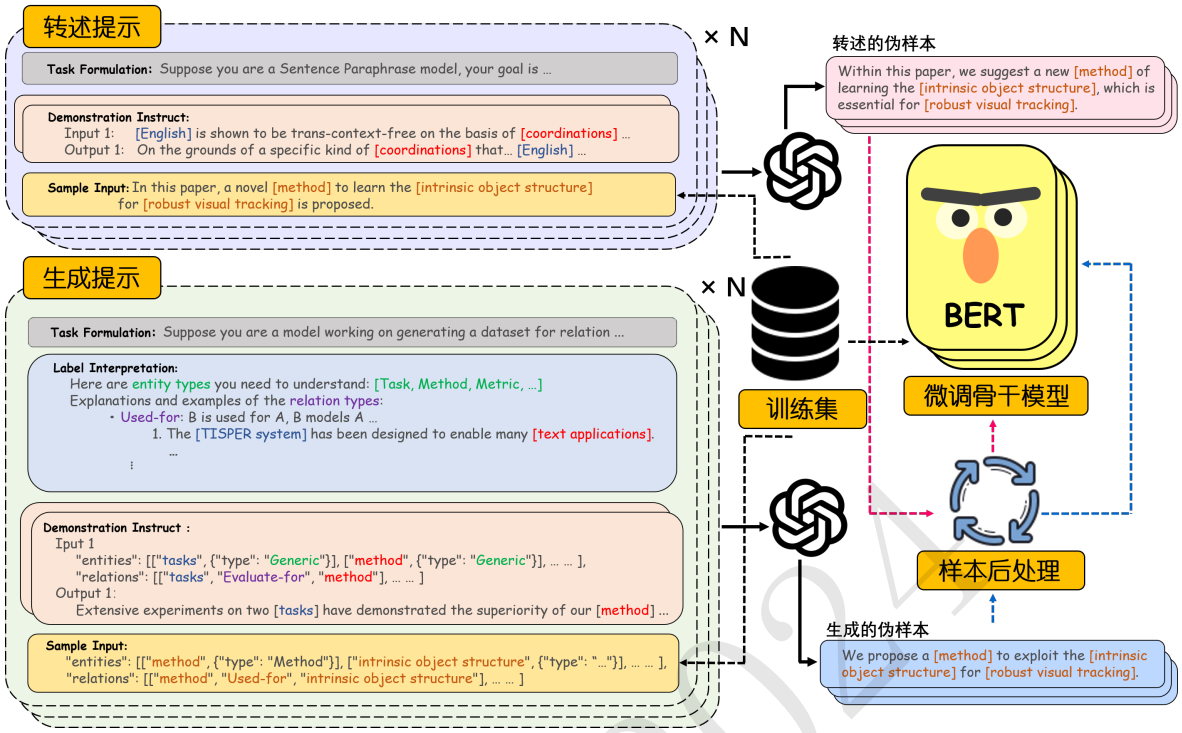



Figure 1: PGA的总体框架。图1左部分是框架的转述式和生成式数据增强方式的提示构成。分别将含有原训练样本的两种提示迭代输入至GPT-3.5中，合成两种伪样本，这些伪样本经过数据后处理，过滤之后转换成各模型需要的格式，连同原有训练样本一起微调各骨干关系抽取模型。

### 3.3 转述(Paraphrase)的数据增强方式

对于每个原训练集的样本$\mathcal{D} = \{(S, e_{s_i}, e_{o_i})\}_{i=1}^{N}$，其中，文本$S$的句意中隐含着关系语义，将$S$和实体跨度的边界信息$\mathcal{E}' = \{e'_i\}_{i=1}^{N}$（不包含实体类型标签）输入至LLM。通过句子转述，在保持原有实体不变的条件下，指导LLM产生一个隐含同样的关系语义但同时与原句子表述、形式不同的新文本$S_P$，最后将其与原有的实体信息连接，得到一个转述产生的伪样本$\mathcal{D}_\mathcal{P} = \{(S_P, e_{s_i}, e_{o_i})\}_{i=1}^{N}$。获得的伪样本$\mathcal{D}_\mathcal{P}$能让RE模型从中学习到比原有样本更多的信息包括同一种关系类型的更多表述和呈现方式，增大正确预测对应关系类型的概率。通过将文本的转述过程视为对字符串$y$的文本序列生成任务，其过程可以写成如下：

$$p_{LLM}\left(S_P|S, e'_s, e'_o\right) = \prod_{l=1}^{L} p\left(y_l|S, e'_s, e'_o, y_{<l}\right) \tag{3}$$

即使用LLM来生成长度为$L$的字符串$y$（转述的句子）的概率。

首先，为每个给定的训练集样本构建转述的数据增强方式的提示，并将其输入至GPT-3.5。如图1左上部分所示，每个转述提示具体由以下3部分组成，完整的转述提示详见附录。

**Task Formulation**：首先命令LLM进行指定的样本转述模型的角色扮演，明确其任务下的身份背景。再提供转述关系抽取样本的任务描述、详细的注意事项，防止LLM转述出不符合规定的样本、增加或者删减实体信息，使LLM进一步明确任务的要求。

**Demonstration Instruct**：为LLM更好学习指定的数据格式和理解转述伪样本的任务，这部分提示为LLM提供了任务中原样本的输入示例和伪样本的输出示例。实践中发现，LLM由

于预训练阶段所学习的先验知识大多是易于人类阅读的形式的文本语料库，其思维方式也更接近于人类的思维方式，所以输入的数据样本中如果出现过多的结构化信息，则会影响LLM对句子的理解。因为基于模型属于跨度级别的RE方法，导致样本全部是JSON格式的数据，嵌套的各类括号、引号，冗长复杂且无用的信息过多，会使LLM将部分样本信息丢失。因此，本文在转述方法的提示中的输入和输出样本格式中，并非使用了数据集原本的易于程序所理解的JSON格式，而是对数据集进行简单改造。将token还原为完整的句子，用方括号将句子中对应实体包裹起来，将实体融合于句中，并不携带实体类型标签($\mathbb{E}$)，仅仅利用实体跨度的边界信息($\mathcal{E}' = \{e'_i\}_{i=1}^{N}$)。这样既能缩短提示中重复的信息，也能便于LLM理解和学习。同时受提示长度限制，我们随机地从原训练集中选择两条样本作为演示，每个示例的输入和输出格式遵循前面提到的数据格式。

**Sample Input**：这部分提示包含一条来源于原训练数据集的样本的的句子和实体跨度边界信息，其作为转述式数据增强任务的输入。每条数据样本需要参照实例中的数据格式，迭代插入转述提示中，后重复输入至LLM中。

### 3.4 生成(Generate)的数据增强方式

将每个原训练集样本$\mathcal{D} = \{(S, e_{s_i}, e_{o_i})\}_{i=1}^{N}$中的句子$S$对应的实体信息$\mathcal{E} = \{e_i\}_{i=1}^{N}$，加上最后模型需预测的结果，即关系标签$\mathcal{R} = \{r_i\}_{i=1}^{N}$，一起输入至LLM。通过LLM强大的文本序列生成能力，得到蕴含全部实体以及关系语义的句子$S_G$。如上文所述，关系抽取的目的是通过模型学习每个训练样本$\mathcal{D}$，预测样本句子中隐含的的关系类型$\mathcal{R}$。而生成的数据增强方式，可以视为这个任务的部分逆序过程——即LLM根据提供的答案标签来生成可学习的样本。对于每个样本，通过其所有已知的标签信息$\mathcal{M} = \{e_{s_i}, e_{o_i}, r_i\}_{i=1}^{N}$，以生成新的句子$S_G$，最后将它与原有的实体信息连接，得到一个生成方式的伪样本$\mathcal{D}_G = \{(S_G, e_{s_i}, e_{o_i})\}_{i=1}^{N}$。这些样本中的生成句子呈现出与原句子不同的描述和表达，暗含完整的标签信息，使模型学习到更多额外的域内标注样本。同样，通过将生成方式的数据增强方法视为对字符串$y$的文本序列生成任务，其过程可以写成如下：

$$p_{LLM}\left(S_G | e_s, e_o, r\right) = \prod_{l=1}^{L} p\left(y_l | e_s, e_o, r, y_{<l}\right) \tag{4}$$

如图1左下部分所示，框架为每个给定的训练集样本构建生成的数据增强方式的提示，提示具体由以下4部分组成，完整的生成提示详见附录。

**Task Formulation**：这部分提示指示LLM扮演RE样本生成模型，规定生成样本的任务定义和任务目标、强调输出的注意事项，防止LLM生成错误样本、改变原实体信息，并且阐述输入文本的格式以及方括号含义，使LLM进一步明确输入输出要求。规定输出文本的体裁和风格，使其更加贴合原数据集句子。由于生成方式输入的实体信息有可能为空，本文规定了没有标签信息的句子经过生成方式获得的伪样本是*"No result can be generated with the given information."*，其包含的实体信息和关系信息为空。

**Demonstration Instruct**：一定数量的输入和输出示例能够帮助LLM理解生成样本的任务，这部分提示为LLM提供了两条随机的样本输入和它们对应的输出。在生成方式的提示中，输出同样采用了方括号包裹句子中的实体的格式。但不同于转述方式的是，生成方式的输入使用格式仍是采用JSON格式，这是因为生成方式的输入不是句子和实体信息，而是完整实体标签信息和关系标签信息。因为输入不包含句子，所以无法使用前面所提出的方括号包裹实体的格式，同时由于JSON格式作为输入的弊端，因此生成方式合成的伪样本错误率会更高。

**Label Interpretation**：数据增强需要确保生成的样本仍然保持实体和关系的标注准确性，保持生成样本的质量和与原始数据尽可能地接近。但同时也要避免生成过于简单和重复的文本，以确保模型能够从中学习到新的信息和模式。由于生成式数据增强任务的特殊性，提示中必须要为LLM提供所有实体标签集合、关系类型标签信息及诠释、数条关系标签的案例以约束LLM，避免其生成的文本的含义与对应的标签信息不符，影响伪样本质量。其中，对应的关系标签描述来源于Luan et al. (2018)。

**Sample Input**：同样，将原训练数据集里的每条样本对应的所有实体和关系信息迭代插入此部分提示中,作为生成方式数据增强任务的输入。

### 3.5 数据后处理

得到LLM输出的句子后，为了保持格式统一，并且减少不必要的API的token消耗，其输出格式同样使用易于阅读的方括号格式。在数据后处理步骤，将得到的句子与其对应的原数据集实体和关系标签信息合并为完整的伪样本，同时与其对应的原样本信息进行比对，识别并且过滤掉错误的伪样本，确保RE模型输入的数据是准确的。最后，将经过筛选的伪样本还原成各RE模型能读取的格式，确保数据的准确性和一致性，保证模型在训练中能够取得更好的效果。

## 4 实验

### 4.1 数据集

本文在一个科学领域关系抽取数据集评估PGA框架。SciERC(Luan et al. (2018))数据集收集了500篇人工智能论文摘要，为其标注了关系和实体信息。我们合成了大量转述和生成的伪样本，以便模型在测试集上评估性能。其合成的两个数据集统计数据如表1所示。

| Dataset | #Train | #Dev | #Test | #Ents(Types) | #Rels(Types) | #Defect Rate |
|---|---|---|---|---|---|---|
| *SciERC* | 1,861 | 275 | 551 | 8,089(6) | 4,716(7) | - |
| $SciERC_P$ | 1,861 | - | - | 5,598(6) | 3,219(7) | 21.60% |
| $SciERC_G$ | 1,589 | - | - | 4,341(6) | 2,402(7) | 14.61% |

Table 1: 原数据集、转述($SciERC_P$)和生成($SciERC_G$)数据集的统计数据

### 4.2 评估指标

对于命名实体识别任务，实体跨度的边界和实体类型需要模型同时正确预测。对于关系抽取任务，本文分别报告了两个评价指标：(1)**边界评估(Rel)**要求模型正确预测每个主体实体和客体实体对的跨度边界和它们之间的关系类型；(2)**严格评估(Rel+)**在Rel的设置下，进一步要求模型对实体的类型也要预测正确(Taillé et al. (2020))。

### 4.3 基线和骨干模型

为了对PGA框架进行评估，本文分别选择以下先进的关系抽取模型作为PGA框架的骨干模型：(1)**SpERT**(Eberts and Ulges (2019))是主流联合实体关系抽取模型，列举文本中所有可能的片段跨度，编码生成跨度的向量表示，并进行跨度的实体标签分类预测。对提取的所有实体跨度两两配对，通过一个多分类模块判断实体对之间的关系类型。(2)**PL-Marker**(Ye et al. (2022))属于流水线模型，为命名实体识别模块提出了面向邻域打包策略，综合考虑了邻域跨度，以更精确地建模实体边界信息。为关系抽取设计了主客实体打包策略，使用悬浮标记和固定标记，让不同客体实体可以共享同一主体实体。(3)**PURE**(Zhong and Chen (2021))是端到端的流水线模型，命名实体识别模块和关系抽取模块独立训练，它们共享相同的预训练编码器。并为关系抽取任务提出悬浮标记，独立处理每对主客实体。还可以选择将其统一放在样本末尾加快训练速度(PURE-Approx)(Zhong and Chen (2021))。此外，本文额外加入了6个主流的关系抽取模型作为基线进行比较。

### 4.4 实施细节

LLM方面，由于截止于进行实验的日期时，OpenAI最先进的基于GPT-3.5长文本处理模型是*text-davinci-003*，同时实践中表明*gpt-3.5-turbo*更适用于聊天对话领域而并非长文本生成领域，所以本文选择*text-davinci-003*作为PGA框架的基干LLM。

在转述式数据增强中，LLM温度设置为0.5，而生成方式里，由于需要更好的样本的多样性，温度被设置为1。本文将原训练集的所有样本进行了转述和生成，在实验中出现了LLM输出错误的样本的情况。因为由转述方式得到伪样本句子拥有对应原样本句子作为清晰的答案对照，所以在数据后处理步骤中，错误的转述样本被过滤筛选出，然后将对应转述错误的样本利用LLM重新合成，直到获得正确的样本为止。而由生成方式合成的伪样本句子对应的原样本输入是标签信息，不便对照，所以错误样本直接被丢弃。两种方式合成伪样本的瑕疵率详

见表1。在将伪样本句子还原成相应模型读取的token格式时，对于那些模型输入需要的但不参与NER和RE任务的无关紧要的样本属性信息（如样本中句子的来源，即ID），我们使用程序统一生成伪属性将其替代。

骨干模型方面，对于SpERT(Eberts and Ulges (2019))，按照原工作，使用*scibert-scivocab-cased*(Beltagy et al. (2019))作为预训练编码器，并遵照设置，使用SciERC的训练集和验证集训练模型；对于PL-Marker(Ye et al. (2022))和PURE(Zhong and Chen (2021))，按照原工作使用的*scibert-scivocab-uncased*(Beltagy et al. (2019))作为预训练编码器，使用SciERC的训练集用以训练模型。并为主要实验和联合伪样本实验中框架下的骨干模型优化模型的超参数，其它实验按照对应原工作的模型参数进行实验，不再优化超参数。

### 4.5 主要结果

将PGA框架主要实验结果与各基线方法和骨干模型性能进行了比较，结果如表2所示。在PGA的两种数据增强方法中，$PGA_P$代表转述方式的数据增强，$PGA_G$代表生成方式的数据增强。从表中可以得知，数据增强对关系抽取来说是一种简单而有效的方法，它可以借用LLM的先验知识来合成伪样本以训练基于微调的模型提升性能：PGA框架基本上取得了Ent，Rel，Rel+三个指标下最高的F1分数，也优于之前的所有基线。在于骨干模型对比方面，基于SpERT(Eberts and Ulges (2019))的$PGA_P$相比原来，在Ent和Rel+的F1分数上取得了最佳，最高提升1.78%，而$PGA_G$相比原来在Rel+上取得1.53%F1分数提升，但在Rel指标下有一定下滑；基于PL-Marker(Ye et al. (2022))的$PGA_P$相比原来，在每个指标上都有一定提升，Rel+的F1分数最高提升2.05%，并且$PGA_G$同样所有指标都有提升，在Ent和Rel的指标上取得所有模型最佳，Rel和Rel+F1分数都提升了1.26%；基于PURE(Zhong and Chen (2021))的$PGA_P$相比原来，同样所有指标都有提升，Rel提升了1.81%，Rel+提升了1.67%，另外PURE的$PGA_G$所有指标也都有提升，Rel+最高提升了1.9%。

同时，我们注意到各个模型转述和生成效果不尽相同，极个别指标的甚至性能会降低。也可以发现，PGA框架在关系抽取方面有主要的提升，尤其是在严格的指标下。

| Model | Encoder | Ent | | | Rel | | | Rel+ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Precision | Recall | **F1** | Precision | Recall | **F1** | Precision | Recall | **F1** |
| SciIE(Luan et al. (2018)) | BiLSTM | 67.20 | 61.50 | 64.20 | 47.60 | 33.50 | 39.30 | - | - | - |
| DyGIE(Luan et al. (2019) | ELMo | - | - | 65.20 | - | - | 41.60 | - | - | - |
| DyGIE++(Wadden et al. (2019)) | $BERT_{Base}$ | - | - | 67.50 | - | - | 48.40 | - | - | - |
| UniRE(Wang et al. (2021)) | | 65.8 | 71.10 | 68.40 | - | - | 40.20† | 37.30 | 36.60 | 36.90 |
| TriMF(Shen et al. (2021)) | | 70.18 | 70.17 | 70.17 | 52.63 | 52.32 | 52.44 | - | - | - |
| PFN(Yan et al. (2021)) | | - | - | 66.80 | - | - | - | - | - | 38.40 |
| SpERT*(Eberts and Ulges (2019)) | | 68.54 | 70.09 | 69.31 | 52.66 | 50.82 | **51.72** | 39.15 | 37.78 | 38.45 |
| SpERT+$PGA_P$ | | 69.43 | 69.67 | **69.55** | 50.72 | 50.51 | 50.62 | 40.31 | 40.14 | **40.23** |
| SpERT+$PGA_G$ | SciBERT | 68.72 | 69.38 | 69.05 | 52.27 | 47.33 | 49.68 | 42.06 | 38.09 | 39.98 |
| PL-Marker(Ye et al. (2022)) | | - | - | 69.90 | - | - | 53.20 | - | - | 41.60 |
| PL-Marker+$PGA_P$ | | 71.15 | 68.78 | 69.95 | - | - | 53.40 | - | - | **43.65** |
| PL-Marker+$PGA_G$ | | 70.11 | 70.86 | **70.48** | - | - | **54.46** | - | - | 42.86 |
| PURE*(Zhong and Chen (2021)) | | 69.33 | 68.55 | 68.93 | 49.69 | 49.69 | 49.69 | 37.37 | 37.37 | 37.37 |
| PURE+$PGA_P$ | | 70.46 | 68.07 | **69.24** | 55.09 | 48.36 | **51.50** | 41.75 | 36.65 | 39.04 |
| PURE+$PGA_G$ | | 70.50 | 67.95 | 69.21 | 55.58 | 44.97 | 49.72 | 43.91 | 35.52 | **39.27** |

Table 2: PGA框架加上各个骨干模型在SciERC测试集上总体的实验结果。性能的评估指标包括查准率、召回率和F1分数。**粗体** 数据表示各骨干模型下的最高性能，带有下划线的数据表示次优的性能。请注意：带有∗符号的模型表示为尽可能获取更多的评估指标，我们根据原作者公布的参数和代码重新运行的实验结果。带有†符号表示引用(Ren et al. (2023))对原模型的复现结果。

## 5 分析

### 5.1 联合两种伪样本

联合两种方式产生的伪样本，与原训练集共同用于微调各骨干模型。将框架联合伪样本$PGA_C$实验结果与各骨干模型性能进行比较，结果如表3所示。在SpERT(Eberts and Ulges


(c) 2024 中国中文信息学会计算语言学专业委员会

(2019))上，$PGA_C$表现不佳，多数性能指标出现不同程度下降；在PURE(Zhong and Chen (2021))上甚至所有指标的性能有都下降；仅在PL-Marker(Ye et al. (2022))上Rel+有0.99%的提升。我们推测是否因为框架合成的伪样本的标注质量存在问题，加之由表1可知，联合起来的伪样本规模远超过原数据集的训练样本规模，其引入大量的实体却未标注、对科研数据集的专有词汇理解不足而产生大量的噪声以致于整个参与训练的数据集的质量低于原训练集的质量，这就导致了对噪声敏感的模型的性能下降。

| Model | Ent | | | Rel | | | Rel+ | | |
|---|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | **F1** | Precision | Recall | **F1** | Precision | Recall | **F1** |
| SpERT*(Eberts and Ulges (2019)) | 68.54 | 70.09 | **69.31** | 52.66 | 50.82 | **51.72** | 39.15 | 37.78 | 38.45 |
| SpERT+$PGA_C$ | 69.40 | 67.72 | 68.55 | 53.00 | 44.46 | 48.35 | 42.72 | 35.83 | **38.97** |
| PL-Marker(Ye et al. (2022)) | - | - | 69.90 | - | - | **53.20** | - | - | 41.60 |
| PL-Marker+$PGA_C$ | 70.61 | 69.44 | **70.02** | - | - | 52.83 | - | - | **42.59** |
| PURE*(Zhong and Chen (2021)) | 69.33 | 68.55 | **68.93** | 49.69 | 49.69 | **49.69** | 37.37 | 37.37 | **37.37** |
| PURE+$PGA_C$ | 69.61 | 67.41 | 68.50 | 49.62 | 47.43 | 48.50 | 37.16 | 35.52 | 36.33 |

Table 3: 在SciERC测试集上联合两种伪样本与原数据集用以微调骨干模型的性能，**粗体** 数据表示各骨干模型下的最高性能。$PGA_C$代表各骨干模型额外地联合了两种数据增强方式合成的伪样本进行训练。

## 5.2 伪样本的质量

根据上面的猜测设计实验，将原数据集从模型的训练里剔除，仅用单独的伪样本以微调各骨干模型。将单独仅含伪样本实验结果与各骨干模型性能进行比较，结果如表4所示。总体上，分别使用单独转述和生成的伪样本进行微调的各模型的性能都出现下降。这些差距表明伪样本的质量与原样本的质量相比还有一定提升的空间。即使在最好情况下，SpERT(Eberts and Ulges (2019))在转述的伪样本单独训练下，Ent的F1分数也下降了3.06%，Rel的F1分数下降了5.88%，Rel+则下降了4.12%。然而，也许是因为PURE(Zhong and Chen (2021))对噪声更加敏感或此实验没有对其超参数进行进一步优化的缘故，在生成的伪样本单独训练中，其F1分数是所有模型中下降得最多的，Ent的F1分数降幅达到了25.56%，Rel的F1分数甚至下降了29.34%，Rel+指标则下降了22.95%。PL-Marker(Ye et al. (2022))的下降幅度则处于两者之间。

总体来说，框架的两种数据增强方式中，转述伪样本质量高于生成伪样本质量。可以发现，转述伪样本微调下的骨干模型的每一个F1分数都超过了对应的生成的伪样本微调下的骨干模型的F1分数，并且每一个指标都相差10%以上。我们考虑造成以上现象的原因主要在于：在实体方面，实体方面的信息主要依靠输入样本提供，而LLM只会在合成样本的时候引入一定量的噪声实体。然而，在关系方面，转述方法产生关系信息几乎只依靠LLM对句子整体含义理解后的关系的推理，因为没有明确的关系信息的输入，LLM需要推导出文本中隐含的关系，这对其识别和推理信息的能力有一定挑战。在产生转述样本的时候，LLM需要生成暗含实体间语义关系的句子，这要求其具备高质量的文本生成能力，以便准确地表达关系信息。此外，生成数据扩增方法产生关系信息的条件更为苛刻，仅输入两种标签信息，没有句子供其参考。这对LLM的条件约束就远远不如转述方式，所以可能在生成句子里LLM就不会引入更隐式的关系语义供模型学习，且其样本还保有一定的噪声，导致其质量不如前者。

## 5.3 伪样本参与微调的数量对性能的影响

分别分割两种方式产生的伪样本，把它们各自以不同数量和原训练集样本一起参与训练，确定是否伪样本数量也会对性能有一定影响，结果如图2所示。可以观察到，对于命名实体识别任务来说，随着转述伪样本的数量的提升，性能先是短暂提升后再次因为过多伪样本加入而下降。然而生成伪样本则经过初始的跌落而后缓慢上升。

另一方面，对于关系抽取任务来说，随着转述伪样本的数量的提升，Rel和Rel+指标的性能分数都在一个特定的数量下达到顶峰，而后随着更多伪样本数量加入，性能逐渐下降，此时其伪样本规模与原训练集规模接近。注意到，之所以命名实体识别和关系抽取的性能拐点不同，可能是因为SpERT(Eberts and Ulges (2019))是一个联合实体关系抽取模型，能一定程度减


(c) 2024 中国中文信息学会计算语言学专业委员会

| Model | Ent | | | Rel | | | Rel+ | | |
|---|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | **F1** | Precision | Recall | **F1** | Precision | Recall | **F1** |
| SpERT*(Eberts and Ulges (2019)) | 68.54 | 70.09 | 69.31 | 52.66 | 50.82 | 51.72 | 39.15 | 37.78 | 38.45 |
| SpERT+$PGA_P^{Sole}$ | 67.24 | 65.28 | 66.25 | 48.67 | 43.33 | 45.84 | 36.45 | 32.44 | 34.33 |
| SpERT+$PGA_G^{Sole}$ | 50.34 | 56.74 | 53.35 | 36.73 | 21.46 | 27.09 | 26.01 | 15.20 | 19.18 |
| PL-Marker(Ye et al. (2022)) | - | - | 69.90 | - | - | 53.20 | - | - | 41.60 |
| PL-Marker+$PGA_P^{Sole}$ | 67.00 | 59.41 | 62.98 | - | - | 44.68 | - | - | 34.63 |
| PL-Marker+$PGA_G^{Sole}$ | 52.67 | 41.61 | 46.49 | - | - | 26.60 | - | - | 19.42 |
| PURE*(Zhong and Chen (2021)) | 69.33 | 68.55 | 68.93 | 49.69 | 49.69 | 49.69 | 37.37 | 37.37 | 37.37 |
| PURE+$PGA_P^{Sole}$ | 59.25 | 55.13 | 57.12 | 40.37 | 31.42 | 35.33 | 29.95 | 23.31 | 26.21 |
| PURE+$PGA_G^{Sole}$ | 52.21 | 37.09 | 43.37 | 34.62 | 14.68 | 20.62 | 24.21 | 10.27 | 14.42 |

Table 4: 在SciERC测试集上仅有单独伪样本参与的骨干模型微调性能。$PGA_P^{Sole}$代表仅有单独转述的伪样本参与训练，$PGA_G^{Sole}$代表仅有单独生成的伪样本参与训练。请注意，SpERT(Eberts and Ulges (2019))原本设置中，原训练集和验证集共同参与了训练。

少误差传递的问题，并且关系抽取总体的性能趋势并未与命名实体识别的性能变化相脱离。这一点在生成伪样本的关系抽取性能上也有所体现。总体上，Rel+性能在更多的伪样本参与下性能再次下降，由此可推测出，命名实体识别和关系抽取的性能并不能随更多伪样本的加入得到持续的提升。

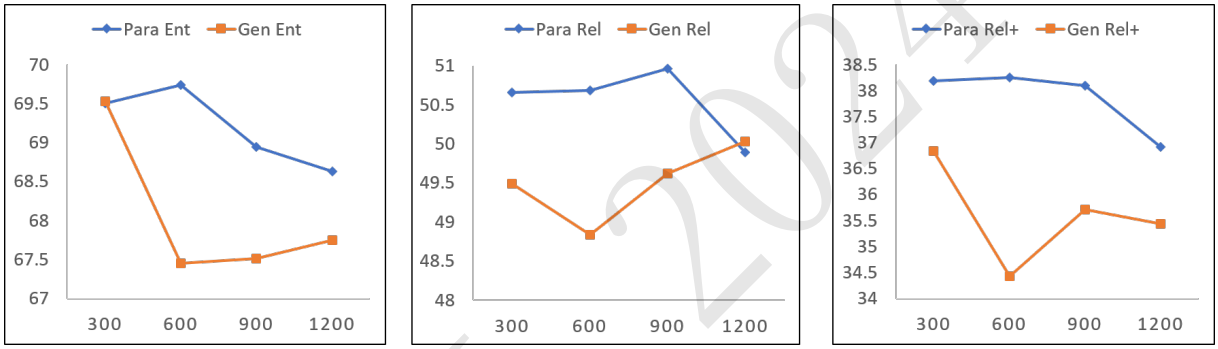



Figure 2: 分别由不同数量伪样本连同原训练集参与训练下，SpERT(Eberts and Ulges (2019))模型的性能变化，横坐标代表参与伪样本数量，纵坐标代表F1分数。

### 5.4 伪样本对原样本的忠实度

为了衡量上面的实验中伪样本具体是否真正接近于真实的原训练集样本，我们使用*sentence-transformers*中的*all-MiniLM-L6-v2*模型，分别对400条转述和生成产生的伪样本，连同它们各自所对应的原训练集样本中的句子进行文本嵌入，再使用t-SNE算法对向量空间中高维的句子嵌入进行降维和可视化处理，得到样本嵌入的空间分布，以观察伪样本对原样本的忠实度。由图3可知，转述的伪样本与原样本保持了相对较高的重叠程度。对于每一个原样本，包括图中边缘零星的样本都有一个非常接近的转述伪样本。然而，生成的伪样本与原样本大致上重叠和贴近，但是从贴近程度来看，其弱于前者。另外从其分布图中右上角可以发现，已经出现了伪样本分布远离对应原样本分布的情况，而转述样本分布未出现此现象。所以，总体上生成样本对原样本的忠实度不及转述样本对原样本的忠实度。

## 6 结论

在这项工作中，我们提出了一个名为PGA的基于大语言模型的数据增强框架，该框架能通过转述和生成两种数据增强方式合成与原数据集对应伪样本，在更具挑战性的科学领域的数据集上，我们对主流的关系抽取模型进行了数据增强，提升了模型的关系抽取的性能。在对产生的伪样本进行具体的分析实验后，本文认为，PGA转述的伪样本在保持对原样本的忠实度较高


(c) 2024 中国中文信息学会计算语言学专业委员会

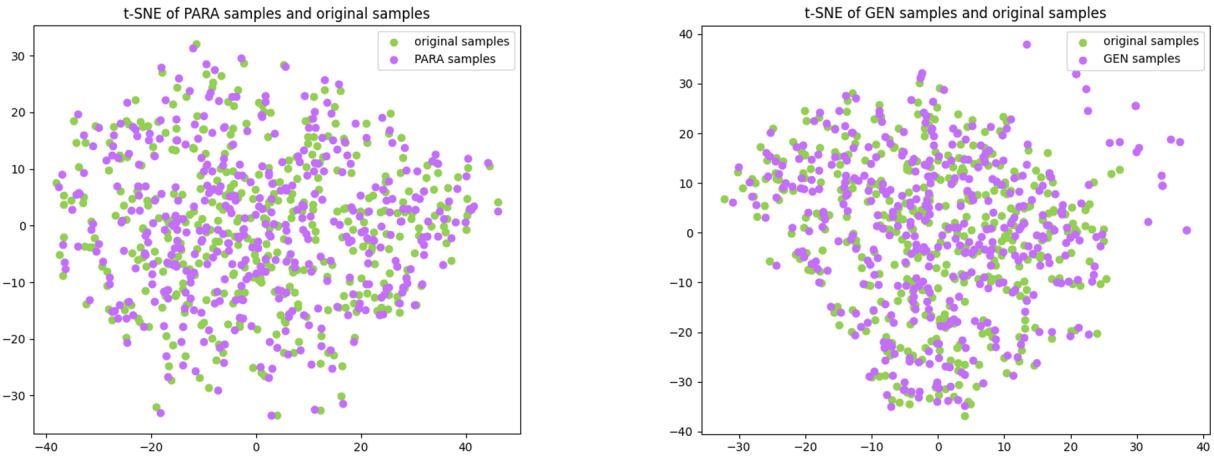



Figure 3: 伪样本与原数据集样本句子的嵌入在向量空间中的分布。

的前提下，能有效引入同种关系的更多语义信息来供模型学习。即使生成的伪样本能提升模型一部分性能，但是其质量还有一定提升的空间。在未来我们需要进一步研究是否可以通过更先进的上下文学习方法或有效的伪数据过滤策略来提升利用大语言模型进行数据增强的精确性和忠实度。

## 致谢

感谢复杂产品智能制造系统技术国家重点实验室开放基金课题《基于多源数据挖掘的复杂产品性能推理与评估方法研究》对本工作提供的资助。

## 参考文献

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China, November. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Tiantian Chen, Lianke Zhou, Nianbin Wang, and Xirui Chen. 2022. Joint entity and relation extraction with position-aware attention and relation embedding. *Applied Soft Computing*, 119:108604.

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Yihan Cao, Zihao Wu, Lin Zhao, Shaochen Xu, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, Hongmin Cai, Lichao Sun, Quanzheng Li, Dinggang Shen, Tianming Liu, and Xiang Li. 2023. Auggpt: Leveraging chatgpt for text data augmentation.


(c) 2024 中国中文信息学会计算语言学专业委员会

Markus Eberts and Adrian Ulges. 2019. Span-based joint entity and relation extraction with transformer pre-training. In *European Conference on Artificial Intelligence*.

Tsu-Jui Fu, Peng-Hsuan Li, and Wei-Yun Ma. 2019. GraphRel: Modeling text as relational graphs for joint entity and relation extraction. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1409–1418, Florence, Italy, July. Association for Computational Linguistics.

Jiahui Gao, Renjie Pi, Yong Lin, Hang Xu, Jiacheng Ye, Zhiyong Wu, Weizhong Zhang, Xiaodan Liang, Zhenguo Li, and Lingpeng Kong. 2023a. Self-guided noise-free data generation for efficient zero-shot learning.

Jun Gao, Huan Zhao, Changlong Yu, and Ruifeng Xu. 2023b. Exploring the feasibility of chatgpt for event extraction.

Ridong Han, Tao Peng, Chaohao Yang, Benyou Wang, Lu Liu, and Xiang Wan. 2023. Is information extraction solved by chatgpt? an analysis of performance, evaluation criteria, robustness and errors.

Tomoki Hayashi, Shinji Watanabe, Yu Zhang, Tomoki Toda, Takaaki Hori, Ramon Astudillo, and Kazuya Takeda. 2018. Back-translation-style data augmentation for end-to-end asr. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 426–433.

Ting-Wei Hsu, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2021. Semantics-preserved data augmentation for aspect-based sentiment analysis. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4417–4422, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Ming Jiang, Jennifer D'Souza, Sören Auer, and J Stephen Downie. 2020. Targeting precision: A hybrid scientific relation extraction pipeline for improved scholarly knowledge organization. *Proceedings of the Association for Information Science and Technology*, 57(1):e303.

Bernal Jimenez Gutierrez, Nikolas McNeal, Clayton Washington, You Chen, Lang Li, Huan Sun, and Yu Su. 2022. Thinking about GPT-3 in-context learning for biomedical IE? think again. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4497–4512, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Wenjun Kong and Yamei Xia. 2023. Care: Co-attention network for joint entity and relation extraction. *arXiv preprint arXiv:2308.12531*.

Ashutosh Kumar, Satwik Bhattamishra, Manik Bhandari, and Partha Talukdar. 2019. Submodular optimization-based diverse paraphrasing and its effectiveness in data augmentation. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3609–3619, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Bo Li, Gexiang Fang, Yang Yang, Quansen Wang, Wei Ye, Wen Zhao, and Shikun Zhang. 2023a. Evaluating chatgpt's information extraction capabilities: An assessment of performance, explainability, calibration, and faithfulness.

Guozheng Li, Peng Wang, and Wenjun Ke. 2023b. Revisiting large language models as zero-shot relation extractors. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6877–6892, Singapore, December. Association for Computational Linguistics.

Guozheng Li, Peng Wang, Wenjun Ke, Yikai Guo, Ke Ji, Ziyu Shang, Jiajun Liu, and Zijie Xu. 2024. Recall, retrieve and reason: Towards better in-context relation extraction.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, Brussels, Belgium, October-November. Association for Computational Linguistics.


(c) 2024 中国中文信息学会计算语言学专业委员会

Yi Luan, Dave Wadden, Luheng He, Amy Shah, Mari Ostendorf, and Hannaneh Hajishirzi. 2019. A general framework for information extraction using dynamic span graphs. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3036–3046, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Xilai Ma, Jing Li, and Min Zhang. 2023a. Chain of thought with explicit evidence reasoning for few-shot relation extraction. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2334–2352, Singapore, December. Association for Computational Linguistics.

Yubo Ma, Yixin Cao, Yong Hong, and Aixin Sun. 2023b. Large language model is not a good few-shot information extractor, but a good reranker for hard samples! In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: Towards zero-shot language understanding.

Makoto Miwa and Mohit Bansal. 2016. End-to-end relation extraction using LSTMs on sequences and tree structures. In Katrin Erk and Noah A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1105–1116, Berlin, Germany, August. Association for Computational Linguistics.

Anders Giovanni Møller, Jacob Aarup Dalsgaard, Arianna Pera, and Luca Maria Aiello. 2024. The parrot dilemma: Human-labeled vs. llm-augmented data in classification tasks.

Minh Van Nguyen, Viet Dac Lai, and Thien Huu Nguyen. 2021. Cross-task instance representation interactions and label dependencies for joint information extraction with graph convolutional networks. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 27–38, Online, June. Association for Computational Linguistics.

Eda Okur, Saurav Sahay, and Lama Nachman. 2022. Data augmentation with paraphrase generation and entity extraction for multimodal dialogue system.

Frédéric Piedboeuf and Philippe Langlais. 2023. Is ChatGPT the ultimate data augmentation algorithm? In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15606–15615, Singapore, December. Association for Computational Linguistics.

Amir Pouran Ben Veyseh, Franck Dernoncourt, Bonan Min, and Thien Nguyen. 2023. Generating labeled data for relation extraction: A meta learning approach with joint GPT-2 training. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11466–11478, Toronto, Canada, July. Association for Computational Linguistics.

Lin Ren, Yongbin Liu, Yixin Cao, and Chunping Ouyang. 2023. CoVariance-based causal debiasing for entity and relation extraction. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2627–2640, Singapore, December. Association for Computational Linguistics.

Tokala Santosh, Prantika Chakraborty, Sudakshina Dutta, Debarshi Sanyal, and Parthapratim Das. 2021. Joint entity and relation extraction from scientific documents: Role of linguistic information and entity types. 09.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In Katrin Erk and Noah A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany, August. Association for Computational Linguistics.

Yongliang Shen, Xinyin Ma, Yechun Tang, and Weiming Lu. 2021. A trigger-sense memory flow framework for joint entity and relation extraction.


(c) 2024 中国中文信息学会计算语言学专业委员会

Changzhi Sun, Yeyun Gong, Yuanbin Wu, Ming Gong, Daxin Jiang, Man Lan, Shiliang Sun, and Nan Duan. 2019. Joint type inference on entities and relations via graph convolutional networks. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1361–1370, Florence, Italy, July. Association for Computational Linguistics.

Bruno Taillé, Vincent Guigue, Geoffrey Scoutheeten, and Patrick Gallinari. 2020. Let's Stop Incorrect Comparisons in End-to-end Relation Extraction! In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3689–3701, Online, November. Association for Computational Linguistics.

Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. 2023. Does synthetic data generation of llms help clinical text mining?

Xuemei Tang, Jun Wang, and Qi Su. 2024. Small language model is a good guide for large language model in chinese entity relation extraction.

Yicheng Tao, Yiqun Wang, and Longju Bai. 2024. Graphical reasoning: Llm-based semi-open relation extraction.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5784–5789, Hong Kong, China, November. Association for Computational Linguistics.

Somin Wadhwa, Silvio Amir, and Byron Wallace. 2023. Revisiting relation extraction in the era of large language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15566–15589, Toronto, Canada, July. Association for Computational Linguistics.

Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying Liu, Haiyue Song, Jiwei Li, and Sadao Kurohashi. 2023. GPT-RE: In-context learning for relation extraction using large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3534–3547, Singapore, December. Association for Computational Linguistics.

Jue Wang and Wei Lu. 2020. Two are better than one: Joint entity and relation extraction with table-sequence encoders. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1706–1721, Online, November. Association for Computational Linguistics.

Yijun Wang, Changzhi Sun, Yuanbin Wu, Hao Zhou, Lei Li, and Junchi Yan. 2021. UniRE: A unified label space for entity relation extraction. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 220–231, Online, August. Association for Computational Linguistics.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China, November. Association for Computational Linguistics.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, Yong Jiang, and Wenjuan Han. 2023. Zero-shot information extraction via chatting with chatgpt.

Benfeng Xu, Quan Wang, Yajuan Lyu, Dai Dai, Yongdong Zhang, and Zhendong Mao. 2023a. S2ynRE: Two-stage self-training with synthetic data for low-resource relation extraction. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8186–8207, Toronto, Canada, July. Association for Computational Linguistics.


(c) 2024 中国中文信息学会计算语言学专业委员会

Canwen Xu, Yichong Xu, Shuohang Wang, Yang Liu, Chenguang Zhu, and Julian McAuley. 2023b. Small models are valuable plug-ins for large language models.

Xin Xu, Yuqi Zhu, Xiaohan Wang, and Ningyu Zhang. 2023c. How to unleash the power of large language models for few-shot relation extraction?

Zhiheng Yan, Chong Zhang, Jinlan Fu, Qi Zhang, and Zhongyu Wei. 2021. A partition filter network for joint entity and relation extraction. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 185–197, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Zhaohui Yan, Songlin Yang, Wei Liu, and Kewei Tu. 2023. Joint entity and relation extraction with span pruning and hypergraph neural networks. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7512–7526, Singapore, December. Association for Computational Linguistics.

Deming Ye, Yankai Lin, Peng Li, and Maosong Sun. 2022. Packed levitated marker for entity and relation extraction. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4904–4917, Dublin, Ireland, May. Association for Computational Linguistics.

Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyoung Park. 2021. GPT3Mix: Leveraging large-scale language models for text augmentation. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2225–2239, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Junjie Yu, Tong Zhu, Wenliang Chen, Wei Zhang, and Min Zhang. 2020. Improving relation extraction with relational paraphrase sentences. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1687–1698, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Chenhan Yuan, Qianqian Xie, and Sophia Ananiadou. 2023. Zero-shot temporal relation extraction with ChatGPT. In Dina Demner-fushman, Sophia Ananiadou, and Kevin Cohen, editors, *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 92–102, Toronto, Canada, July. Association for Computational Linguistics.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. An autoregressive text-to-graph framework for joint entity and relation extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19477–19487.

Chenglong Zhang, Shuyong Gao, Haofen Wang, and Wenqiang Zhang. 2022. Position-aware joint entity and relation extraction with attention mechanism. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4496–4502. International Joint Conferences on Artificial Intelligence Organization, 7. Main Track.

Kai Zhang, Bernal Jimenez Gutierrez, and Yu Su. 2023. Aligning instruction tasks unlocks large language models as zero-shot relation extractors. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 794–812, Toronto, Canada, July. Association for Computational Linguistics.

Zexuan Zhong and Danqi Chen. 2021. A frustratingly easy approach for entity and relation extraction. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 50–61, Online, June. Association for Computational Linguistics.

Yuqi Zhu, Xiaohan Wang, Jing Chen, Shuofei Qiao, Yixin Ou, Yunzhi Yao, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2024. Llms for knowledge graph construction and reasoning: Recent capabilities and future opportunities.

CCL 2024

# 附录

| Paraphrase Prompt | |
|---|---|
| **Task Formulation** | Suppose you are a sentence paraphrase model. Your goal is to follow a given sentence and the entities in the sentence, describe the sentence in different words than the original sentence while keeping the entities in the sentence and the meaning of the sentence intact, then output it.<br>**Note:** In the sentence to be processed, the part enclosed in parentheses is the entity; do not add, delete, or change the entity, or output any redundant information. Please fully understand the meaning of the sentence before paraphrasing. |
| **Demonstration Instruct** | Please refer to the following examples for the format of the output, and please study the following examples to accomplish the above tasks more accurately.<br><br>**Iput1:** '[English] is shown to be trans-context-free on the basis of [coordinations] of the respectively type that involve [strictly syntactic cross-serial agreement].'<br>**Output1:** On the grounds of a specific kind of [coordinations] that require [strictly syntactic cross-serial agreement], [English] is demonstrated to be trans-context-free.<br><br>**Iput2:** 'The [agreement] in question involves number in [nouns] and [reflexive pronouns] and is syntactic rather than semantic in nature because [grammatical number] in [English], like [grammatical gender] in [languages] such as [French] , is partly arbitrary.'<br>**Output2:** The number in [nouns] and [reflexive pronouns] that must match in this [agreement] is based on syntax rather than meaning, since [English] [grammatical number], like [grammatical gender] in [languages] like [French], is partly random. |
| **Sample Input** | **Sentence:** In this paper we show how two standard [outputs] from [information extraction (IE) systems]-[named entity annotations] and [scenario templates]-can be used to enhance access to [text collections] via a standard [text browser]. |

Table 5: PGA转述的数据增强方法的详细提示

CCL 2024

| Generate Prompt | |
|---|---|
| **Task Formulation** | Suppose you are a model working on generating a dataset for relation extraction. One data sample of a relation extraction dataset is a sentence, and its corresponding entity information and relation information. Entities are embedded in the sentence, entities definitely have entity types, and certain pairs of entities will also contain corresponding relation types based on semantics. **Your goal:** to generate sentences that contain information about entities and relations, given that information. |
| **Label Interpretation** | Here is the information you need to understand: 6 entity types: **[Task, Method, Metric, Material, Generic, OtherScientificTerm]**<br>Below are explanations and examples of the relation types, with pairs of entities where a relation exists enclosed in parentheses.<br>• **Used-for:** B is used for A, B models A, A is trained on B, B exploits A, A is based on B.E.g.<br>1. The [TISPER system] has been designed to enable many [text applications].<br>2. Our [method] models [user proficiency].<br>3. Our [algorithms] exploits [local soothness].<br>• **Feature-of:** B belongs to A, B is a feature of A, B is under A domain. E.g.<br>1. [prior knowledge] of the [model]<br>2. [genre-specific regularities] of [discourse structure]<br>3. [English text] in [science domain]<br>• **Hyponym-of:** B is a hyponym of A, B is a type of A.E.g.<br>1. [TUIT] is a [software library]<br>2. [NLP applications] such as [machine translation] and [language generation]<br>• **Part-of:** B is a part of A... E.g.<br>1. The [system] includes two models: [speech recognition] and [natural language understanding]<br>2. We incorporate [NLU module] to the [system].<br>• **Evaluate-for:** B is evaluated for A, A use B to get estimation, E.g.<br>1. [Intra-sentential quality] is evaluated with [rule-based heuristics]<br>2. We describe a new [system] that enhances Criterion's capability , by evaluating multiple aspects of [coherence in essays].<br>• **Compare:** Opposite of conjunction, compare two models/methods, or listing two opposing entities. E.g.<br>1. Unlike the [quantitative prior], the [qualitative prior] is often ignored...<br>2. We compare our [system] with previous [sequential tagging systems]...<br>• **Conjunction:** Function as similar role or use/incorporate with. E.g.<br>1. obtained from [human expert] or [knowledge base]<br>2. NLP applications such as [machine translation] and [language generation] |

CCL 2024

续表

| Generate Prompt | |
|---|---|
| **Demonstration Instruct** | **Caution:** Do not leave out any entities, if the 'entities' is empty, output 'No result can be generated with the given information.', and please try to avoid generating entities in the output sentence that are not provided by the input, such as 'Problem', 'Task', 'Method' and so on. Do not leave out any relation or add relation between pairs of entities that don't have given relation. The genre of each sample sentence generated is the abstract of an AI paper. Please study the following examples and refer to them for the output format (entities enclosed in square brackets in the output sentences)<br><br>**Input1:{'entities':** [['agreement', {'type': 'Generic'}], ['nouns',{'type': 'OtherScientificTerm'}], ['reflexive pronouns', {'type': 'OtherScientificTerm'}], ['grammatical number', {'type': 'OtherScientificTerm'}], ['English', {'type': 'Material'}], ['grammatical gender', {'type': 'OtherScientificTerm'}], ['languages', {'type': 'Material'}], ['French', {'type': 'Material'}]], **'relations':** [['nouns', 'Conjunction', 'reflexive pronouns'], ['grammatical gender', 'Feature-of', 'languages'], ['French', 'Hyponym-of', 'languages']]}<br>**Output1**: 'The [agreement] in question involves number in [nouns] and [reflexive pronouns] and is syntactic rather than semantic in nature because [grammatical number] in [English], like [grammatical gender] in [languages] such as [French] , is partly arbitrary.'<br><br>**Input2:{'entities':** [['tasks', {'type': 'Generic'}], ['method', {'type': 'Generic'}], ['state-of-the-art methods', {'type': 'Generic'}]], **'relations':** [['tasks', 'Evaluate-for', 'method'], ['method', 'Compare', 'state-of-the-art methods']]}<br>**Output2**: 'Extensive experiments on two [tasks] have demonstrated the superiority of our [method] over the [state-of-the-art methods].' |
| **Sample Input** | Now please generate the output based on my input.<br>**Input: {'entities':** [['method', {'type': 'Method'}], ['intrinsic object structure', {'type': 'OtherScientificTerm'}], ['robust visual tracking', {'type': 'Task'}]], **'relations':** [['method', 'Used-for', 'intrinsic object structure'], ['intrinsic object structure', 'Used-for', 'robust visual tracking']]} |

Table 6: PGA生成的数据增强方法的详细提示

第二十三届中国计算语言学大会论文集，第352页-第369页，太原，中国，2024年7月25日至28日。
卷1：主会论文
(c) 2024 中国中文信息学会计算语言学专业委员会

# UFSC：基于统一特征空间构建的零样本关系抽取

**刘雨辰[1,2*]，段建勇[1,2*]，孙康[3]，张晴[1,2]，何丽[1,2]，王昊[1,2]，刘杰[1,2*]**
[1]北方工业大学信息学院，北京100144
[2]CNONIX 国家标准应用与推广实验室，北京100144
[3]江苏自动化研究所，江苏，连云港222006
1870806242@mail.ncut.edu.cn; duanjy@ncut.edu.cn; liujxxxy@126.com

## 摘要

零样本关系抽取（ZSRE）旨在从可见关系中学习提取不可见关系的能力。一些研究表明：将样本语句与关系描述匹配进而预测不可见关系的方法，可以有效完成零样本关系抽取任务。然而，现有的匹配框架方法很少统一样本语句与关系描述的特征空间，缺乏对二者特征进行对齐。因此，本文提出一种为匹配框架零样本关系抽取而设计的统一特征空间构建方法。统一样本语句与关系描述的编码模块，并在此基础上引入特征相似损失。同时，为了减轻特征在空间上的聚合现象，引入特征均匀化模块，旨在构建特征更加均匀化的特征空间。本文所提出的方法实现了性能上的提升。与之前最佳的结果相比，在FewRel和Wiki-ZSL数据集上$F_1$值平均提高1.6%和3.4%，体现了统一特征空间构建以及特征均匀化模块的有效性。

**关键词：** 关系抽取；零样本；特征空间

# UFSC: A Unified Feature Space Construction for Zero-Shot Relation Extraction

**Yuchen Liu[1,2*], Jianyong Duan[1,2*], Kang Sun[3], Qing Zhang[1,2], Li He[1,2], Hao Wang[1,2], Jie Liu[1,2*]**
[1]School of Information Science and Technology, North China University of Technology, Beijing 100144,China
[2]CNONIX National Standard Application and Promotion Lab, Beijing 100144,China
[3]Jiangsu Automation Research Institute of CSIC, Lianyungang 222006
1870806242@mail.ncut.edu.cn; duanjy@ncut.edu.cn; liujxxxy@126.com

## Abstract

Zero-shot relation extraction (ZSRE) seeks to develop the capability to extract hidden relations from observable ones. Previous studies have highlighted that methods matching sample statements with relation descriptions to predict hidden relations can effectively tackle the zero-shot relation extraction task. Nevertheless, current matching framework approaches often fail to integrate the feature space of sample statements and relation descriptions, thus lacking feature alignment between the two. Therefore, this paper introduces a unified feature space construction method tailored for zero-shot relation extraction within a matching framework. We integrate the encoding modules of sample statements and relation descriptions, supplemented with a feature similarity loss. Additionally, to address feature spatial aggregation issues, we introduce a feature homogenization module to foster a more consistent feature space. The method introduced in this paper achieves a performance improvement. The $F_1$ values on the FewRel and Wiki-ZSL datasets are improved by an average of 1.6% and 3.4% compared to the previous best results, demonstrating the effectiveness of the unified feature space construction as well as the feature homogenization module.

第二十三届中国计算语言学大会论文集，第370页-第381页，太原，中国，2024年7月25日至28日。
卷1：主会论文
(c) 2024 中国中文信息学会计算语言学专业委员会

**Keywords:** Relation Extraction , Zero-Shot , Feature Space

## 1 引言

关系抽取（RE）是自然语言处理（NLP）的一项基本任务，目的在于识别文本当中一对实体之间的语义关系。在高质量标记的数据上学习时，关系抽取模型取得了较好的性能。然而，新兴关系类型不断出现，无休止地标注数据成本较高也不切实际。

为了处理缺乏标注数据的新兴关系类型，零样本关系抽取（ZeroRE）最近受到较多关注。两项研究 (Chen and Li, 2021; Zhao et al., 2023)使用固定的Sentence-BERT编码器 (Reimers and Gurevych, 2019)对关系描述进行向量化表示，在零样本条件下利用匹配式模型完成关系抽取任务。两项研究均取得较大的性能提升。

但是，固定的关系描述编码模块也存在着一些局限性。首先，模型只能对样本语句编码器进行训练，使其生成的特征尽可能拟合描述编码器的特征空间。在此过程中，由于描述特征不可学习的特点，模型无法从描述中学习到较多的特征信息，对其在特征空间上的表示并不充分。而零样本任务下的已知信息较少，信息中含有的特征较为珍贵。特征不充分的表达，是低资源场景下的一种资源浪费。

同时，固定的关系描述编码模块对特征空间的构建也不完善，不利于模型区分特征相近的关系。如Table 1所示，关系标签“Country”与“Country of citizens”，二者的关系标签、描述都较为接近。在描述中，虽然“Country”标识该关系不能作用于人类，但是二者含有“state”“country”等较为相似的词汇，不利于模型对二者的关系描述特征进行区分。并且，二者之间的语义相似度也较高，模型对于关系标签“Country”与“Country of citizens”之中的语义特征也较易混淆。

| 关系标签 | 关系描述 |
|---|---|
| Country | country is sovereign **state** of this item; don't use on humans |
| Country of citizens | country of citizenship is the object is a **country** that recognizes the subject as its citizen |

Table 1: 相似关系标签示例。

在特征空间上，相似度更高的文本信息，特征更为临近。而在预测阶段，对于样本语句与关系描述的匹配，也是通过计算样本与每个描述特征之间的相邻度完成。若两组描述的特征临近，模型难以分辨样本应该与哪一组描述匹配。因此，描述相邻度越高，越易发生样本特征的匹配错误。

基于以上问题，本文取消此前固定的关系描述编码模块，将其设置为可学习状态，使得模型在训练过程中可以充分学习到描述中词汇、语义等特征，尽可能利用零样本场景下的已知信息。并且，在此基础上统一样本语句与关系描述的特征空间，使二者的特征匹配处于相同的特征空间进行，提升模型匹配框架性能。同时，提出一种特征空间构建方法，使得模型对于特征空间上的相似性判断更加准确，空间上的特征分布也更加均匀。具体而言，引入特征相似损失替代先前的交叉熵损失，将训练阶段中对于关系标签的多分类任务转化为对于样本特征与描述特征相似度判断的二分类任务，更加拟合预测阶段模型对关系预测的过程。同时，针对特征聚合现象，设置特征均匀化模块。该模块将特征空间上较为临近的特征进行分散，使得一个样本语句周围较少出现两个甚至多个相邻的关系描述特征，进而提升模型在匹配过程中的准确度。

如Figure 1所示，模型UFSC在相似关系标签“Country”与“Country of citizens”上的抽取性能，较RE-Matching (Zhao et al., 2023)有较大提升。错误样本如Table 2所示，UFSC在“Country”标签中的FN错误样本相较于RE-Matching有大幅减少，即标签

*通讯作者
该工作受到国家科技部科技创新2030-“新一代人工智能”重大项目(2020AAA0109700)，国家自然科学基金(61972003,62076167) 资助。
©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

为“Country”样本被识别为其他标签的情况有所减少。同时，UFSC在“Country of citizens”标签中的FP错误样本有所减少，即其他标签匹配为“Country of citizens”的情况有所减少。这得益于UFSC中的描述特征均匀化方法可以对关系描述的特征有区别性更强特征表示。并且，训练阶段负样本检索方法挑选出的负样本迷惑性更强，UFSC在预测阶段对于负样本的识别率更高。

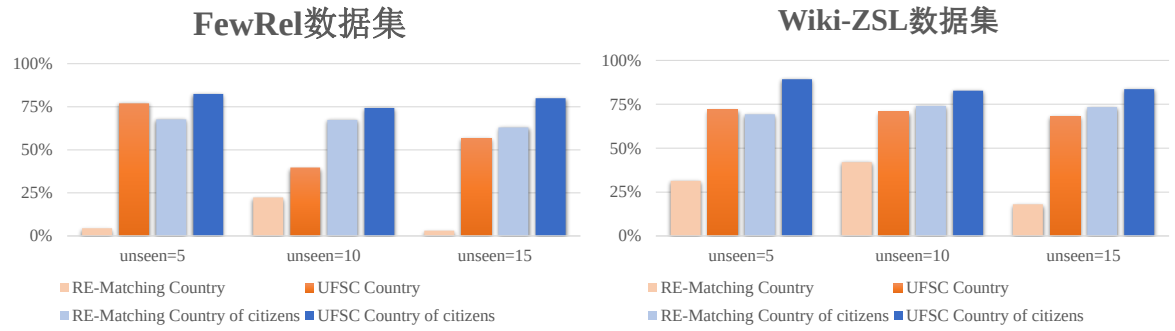



Figure 1: 相似关系标签的$F_1$值对比。

| 数据集 | 不可见关系数量 | 关系标签 | RE-Matching | | UFSC | |
|---|---|---|---|---|---|---|
| | | | FP | FN | FP | FN |
| FewReL | *unseen*=5 | Country | 5 | 684 | 12 | **254** |
| | | Country of citizens | 611 | 5 | **233** | 45 |
| | *unseen*=10 | Country | 104 | 599 | 10 | 524 |
| | | Country of citizens | 592 | 44 | 454 | 18 |
| | *unseen*=15 | Country | 5 | 689 | 19 | **415** |
| | | Country of citizens | 734 | 41 | **320** | 20 |
| Wiki-ZSL | *unseen*=5 | Country | 170 | 781 | 171 | **342** |
| | | Country of citizens | 854 | 16 | **232** | 7 |
| | *unseen*=10 | Country | 84 | 711 | 191 | **348** |
| | | Country of citizens | 658 | 24 | **397** | 15 |
| | *unseen*=15 | Country | 73 | 893 | 228 | **362** |
| | | Country of citizens | 662 | 37 | **317** | 54 |

Table 2: 相似关系标签的错误样本数量分析。

## 2 相关工作

### 2.1 关系抽取

关系抽取旨在根据相关上下文提取两个给定实体之间的关系，其在信息抽取和知识图谱的构建中发挥着关键作用。早期的方法涉及基于CNN、RNN (Zhang et al., 2017)和基于图 (Guo et al., 2021)的方法。随着预训练语言模型的发展 (Devlin et al., 2018)，使用预训练模型进行关系抽取任务已经成为一种主流范式。

### 2.2 零样本关系抽取

零样本关系抽取旨在识别出没有经过训练的新关系。Levy等人 (2017)将零样本关系抽取任务转化为问答形式。预先定义多个问题模板，将关系设置为问题答案，通过找到最符合给定问题的答案完成关系抽取任务。Obamuyide和Vlachos (2018)等人训练一个文本蕴含模型，来预测输入的关系描述是否可以从文本中推导出来。再通过训练阶段可见的语义关系推广到预测阶段不可见的语义关系。Chen和Li (2021)等人设计一个多任务学习模型，通过学习最小化样本与关系描述之间的距离，对关系进行分类。具体而言，将样本和关系描述映射到嵌入空间，利用最


(c) 2024 中国中文信息学会计算语言学专业委员会

邻近搜索来预测不可见的关系。Zhao和Zhan (2023)等人提出一种细粒度匹配方法，将句子间匹配分解为头尾实体匹配和上下文匹配。同时随机采样关系描述作为负样本对，基于匹配分数对不可见关系进行预测。

与上述方法相比，本文的方法有三个显著优势：

1. 统一样本语句与关系描述的特征空间，将先前样本语句特征单方向拟合关系描述特征空间的任务，变为构建一个样本语句与关系描述共享的特征空间。

2. 引入特征相似损失，使得模型在训练阶段学习判断相似特征的能力，进而提升预测阶段的泛化效果。

3. 模型对相似性较高的关系区分能力较强。通过设置特征均匀化模块，对特征空间上聚合度较高的特征进行分散。

## 3 方法

本节提出一种关系描述重构策略，丰富其中的可学习特征。同时，提出一种为匹配框架零样本关系抽取而设计的统一特征空间构建方法。通过设置可学习的关系描述编码模块将其与样本语句共享，并引入特征相似损失，构建统一的特征空间。提出一种特征均匀化方法。通过分散相似度较高的关系描述特征，并在负样本随机采样的基础上融合负样本检索，帮助模型推离错误匹配的描述特征，以得到更为均匀化的特征空间。

### 3.1 问题定义

零样本关系抽取（ZeroRE）任务的目标，是从可见关系集合$R_s = \{r_1^s, r_2^s, \ldots, r_n^s\}$学习到对应知识，并推理至不可见关系集合$R_u = \{r_1^u, r_2^u, \ldots, r_n^u\}$中。关系集合$R_s$和$R_u$是不相交的，即$R_s \cap R_u = \emptyset$。并且在训练阶段只有可见关系集合$R_s$中的样本可以使用。本文将零样本关系抽取转化为特征匹配任务。在训练阶段，给定含有$n$个样本的训练集$D_s = \{(t_i, t_i^{e1}, t_i^{e2}, r_i^s, d_i)|i = 1, \ldots, n\}$，由样本语句$t_i$、头实体$t_i^{e1}$、尾实体$t_i^{e2}$、可见关系标签$r_i^s$及其对应的关系描述$d_i$组成。模型$M$在训练集上对可见关系$r_s$所对应的样本$t_i$与描述$d_i$进行特征匹配，学习统一的编码模块$E$，以得到更加准确的样本和描述的特征表示。在预测阶段，将含有$m$个样本的测试集$D_u = \{(t_i, e_{1i}, e_{2i})|i = 1, \ldots, m\}$和k个描述$d = \{d_1, d_2, \ldots, d_k\}$输入至模型$M$中，找到与样本$t_i$特征相似度最高的描述$d_i$，将其对应关系$r_i^u$作为该样本的关系预测结果，组成关系三元组$(t_i^{e1}, t_i^{e2}, r_i^u)$。模型整体结构如Figure 2所示。

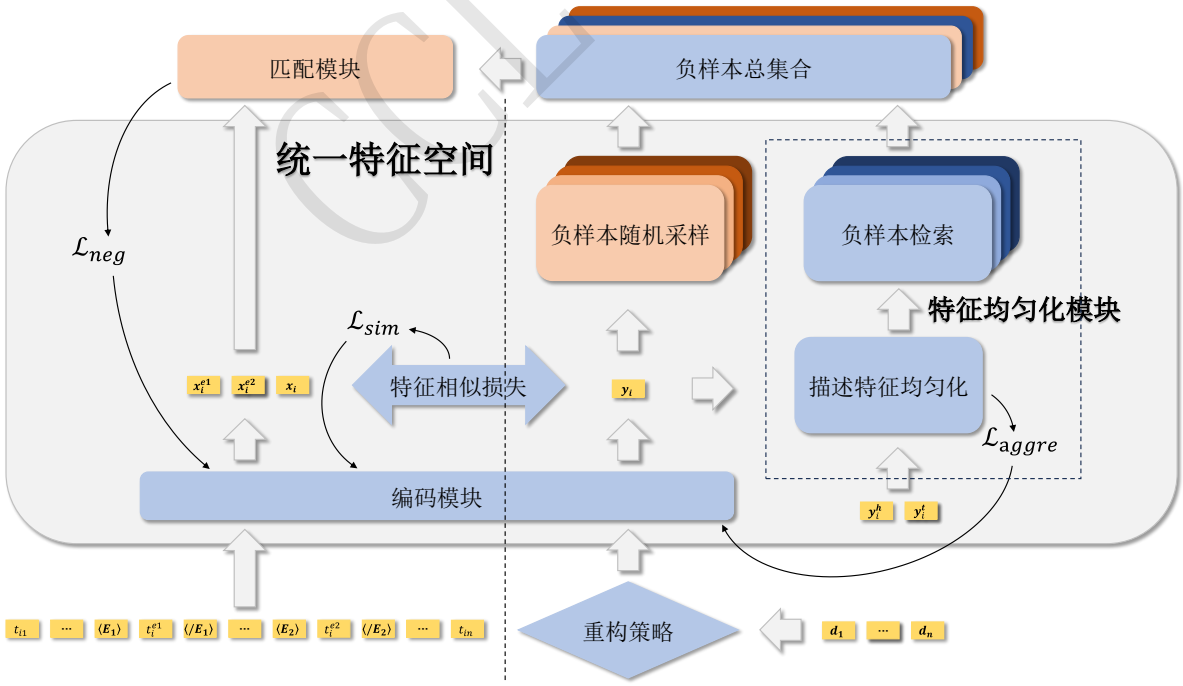



Figure 2: UFSC模型结构图。


(c) 2024 中国中文信息学会计算语言学专业委员会

### 3.2 关系描述重构策略

为了得到信息更为丰富的特征，模型对关系描述进行重新构建。将描述转化为提示与关系定义相结合的方式，在现有的基础上融合关系作用的信息。首先，设置提示模板，采用“the relation between head entity and tail entity is $R_i$”的形式提示模型，头实体与尾实体之间的关系标签为$R_i$。之后，将标签与定义组合，采用“$R_i$ is $D_i$”的形式，对标签的语义信息作进一步解释。

例如，对于标签“Mother”，给出新的描述“the relation between head entity and tail entity is mother, mother is female parent of the subject”。提示模板对描述的特征信息进行补充，在原有描述的基础上，加入关系作用，丰富描述的特征信息。定义对标签进行更深层次的解释，在特征空间上得到更加准确的描述特征，提高关系描述与样本语句的同义程度，使得二者的特征在特征空间上的相似度更高。

但是，丰富的特征在可学习样本较少的场景下，更易产生噪声。因此，模型描述重构策略为：只在不可见关系较少的场景下加入提示模板；对于不可见关系较多的场景，仅使用定义作为描述。

### 3.3 统一特征空间构建

模型将关系描述编码模块与样本语句编码模块统一，并引入特征相似损失，以帮助编码模块对关系标签相同的样本与描述，给出相似性更高的特征表示。

#### 3.3.1 编码模块

为了尽可能利用关系描述中的特征信息，模型设置可学习的关系描述编码模块。假设输入关系描述$d_i = [d_{i1}, d_{i2}, \ldots, d_{in}]$包含$n$个词元。沿用Zhao和Zhan等人 (2023)的方法，对描述中的头尾实体进行特征表示，得到与$y_i^h$与$y_i^t$。然后使用$Bert$ (Devlin et al., 2018)的$pooler_output$层来获得描述的特征表示$y_i$。

$$y_i = Bert_{encoder}(d_i) \tag{1}$$

统一的特征空间，更易于对样本与描述进行匹配。将描述编码模块与样本共享。假设输入文本$t_i = [t_{i1}, t_{i2}, \ldots, t_{in}]$包含$n$个词元，$t_i^{e1}$和$t_i^{e2}$分别表示头实体和尾实体。首先，为了更好地从样本中提取实体特征信息，用实体特殊标记$<E_1>$、$</E_1>$、$<E_2>$、$</E_2>$分别标记$t_i^{e1}$和$t_i^{e2}$的起始和结束的位置 (Soares et al., 2019)。然后使用同一编码模块$Bert$来获得样本的隐藏层表示$h_i^t$，同时使用$Maxpool$和激活层tanh来获得头尾实体特征$x_i^{e1}$、$x_i^{e2}$。

$$h_{i1}^t, \ldots, h_{in}^t = Bert_{encoder}(t_{i1}, \ldots, <E_1>, t_i^{e1}, </E_1>, \ldots, <E_2>, t_i^{e2}, </E_2>, \ldots, t_{in}) \tag{2}$$

$$x_i^e = \tanh(Maxpool(h_{<E>}, h_i^e, h_{</E>})) \tag{3}$$

与描述编码模块相同，将$Bert$的$pooler_output$层输出经过激活层后，作为样本的特征表示$x_i$。

$$x_i = \tanh(Bert_{encoder}(t_i)) \tag{4}$$

#### 3.3.2 特征相似损失

为了更加拟合关系预测过程，模型引入特征相似损失$\mathcal{L}_{sim}$。计算关系描述$y_i$与样本语句$x_i$在特征空间中的相邻度$\cos(x_i, y_i)$，将其作为特征相似性$Sim(x_i, y_i)$。如果二者的特征相邻度越高，即特征在特征空间中夹角越小，就认为特征之间的相似性越高。

$$Sim(x_i, y_i) = \cos(x_i, y_i) \tag{5}$$

参照模型在预测阶段的匹配模式，设置特征相似损失$\mathcal{L}_{sim}$。具体而言，希望模型可以学习到关系标签相同的样本与描述之间更多的特征相似性。最理想的情况，标签相同的样本特征与描述特征完全重合，即相似性$Sim$取得相邻度最大值1。因此，将二者特征相似性向最大值1拟合，即计算特征相似性与最大相似性1之间的均方误差$MSE$，作为特征相似损失$\mathcal{L}_{sim}$。如此模

CCL 2024

型$M$在训练过程中，通过不断减少均方误差值，优化其共享的编码模块。在预测阶段，编码模块对相同标签的样本与描述在特征空间上的特征表示，也更为相邻。由此提升模型的匹配准确率。

$$\mathcal{L}_{sim} = MSE(Sim, 1) \tag{6}$$

### 3.4 特征均匀化模块

如Figure 3所示，若描述特征$y_i$与$y_j$在特征空间中聚合度较高，与$y_i$标签相同的样本特征$x_{i1}$、$x_{i2}$在匹配预测过程中，可能会被模型错误认为与$y_j$相匹配。如果仅使用负样本对比机制，作为负样本的描述特征$y_j$可能只会在$x_{i1}$与$x_{i2}$之间反复波动，不跳出该范围而远离正样本$y_i$。针对以上情况，模型设置特征均匀化模块，由描述特征均匀化与负样本检索两部分组成，帮助模型区分聚合较高的特征。

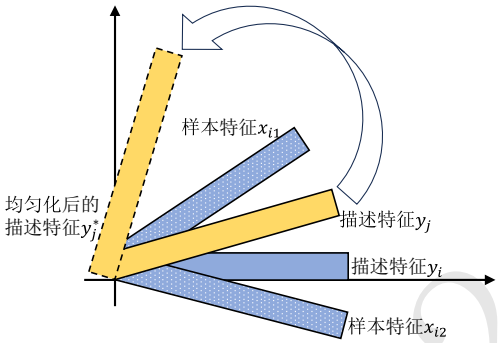



Figure 3: 描述特征聚合现象。

#### 3.4.1 描述特征均匀化

模型通过计算特征夹角的方式，得到描述特征空间附近的特征聚合度$Aggre$。

$$Aggre_{ij}(y_i, y_j) = (1 - 2 \cdot \alpha) \cdot \cos(y_i, y_j) + \alpha \cdot \cos(y_i^h, y_j^h) + \alpha \cdot \cos(y_i^t, y_j^t) \tag{7}$$

例如，对于关系描述$y_i$，将其余所有关系描述$y_{aggre} = \{y_1, y_2, \ldots, y_k | y_i \notin y\}$看作是易聚合描述集合。计算$y_i$与集合中每个负样本之间的特征聚合度$Aggre_i = \{Aggre_{i1}(y_i, y_1), Aggre_{i2}(y_i, y_2), \ldots, Aggre_{ik}(y_i, y_k)\}$。之后，从聚合度集合$Aggre_i$中找到与$y_i$聚合度最高的特征$y_{max}$。将其聚合度与最高聚合限度$margin$进行比较。若其值小于$margin$，则表示$y_{max}$与$y_i$在特征空间聚合度较高，在之后匹配过程中易出现混淆现象。同时设置描述特征均匀损失$\mathcal{L}_{aggre}$。

$$\mathcal{L}_{aggre} = \max(Aggre_{imax}(y_i, y_{max}) - margin, 0) \tag{8}$$

#### 3.4.2 负样本检索

先前任务的负样本采样机制为随机采样 (Zhao et al., 2023)。由于随机性的存在，模型并不能准确挑选出对于样本特征$x_i$最具迷惑性的描述特征$y_j$，如Figure 4所示。因此，若仅沿用负样本随机采样机制，匹配过程中仍易发生样本特征与错误描述特征相互匹配的现象。对此，模型在负样本随机采样的基础上融合负样本检索。

在原先含有$ra$个负样本随机采样集合$y_{random}$中加入负样本检索集合$y_{retriever}$，共同组成负样本总集合$y_{neg}$。其中，集合$y_{retriever}$由与描述特征$y_i$聚合度最高的$re$个描述特征组成。

$$y_{neg} = y_{random} \cup y_{retriever} \tag{9}$$

沿用Zhao和Zhan等人 (2023)的方法，使用细粒度匹配机制计算样本特征$x_i$与负样本集合$y_{neg}$中每个负样本的相似分数$Score_{ij}(x_i, y_j)$，并计算负样本损失$\mathcal{L}_{neg}$。其中$interval$为超参数，表示样本特征与描述特征之间的最小相邻间隔。

CCL 2024

$$Score_{ij}(x_i, y_j) = (1 - 2 \cdot \beta) \cdot \cos(x_i, y_j) + \beta \cdot \cos(x_i^{e1}, y_j^h) + \beta \cdot \cos(x_i^{e2}, y_j^t) \tag{10}$$

$$\mathcal{L}_{neg} = \max(Score_{ii}(x_i, y_i) - Score_{imax}(x_i, y_{max}) + interval, 0) \tag{11}$$

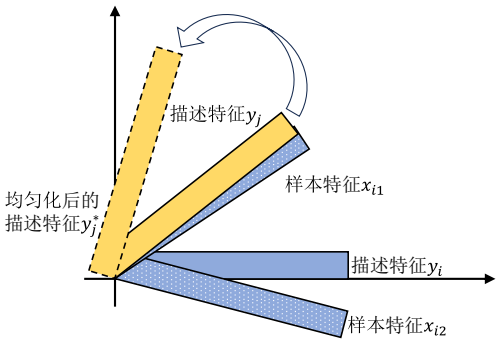



Figure 4: 样本描述特征聚合现象。

### 3.5 模型训练与预测

在训练阶段，模型按先后顺序分别对$\mathcal{L}_{sim}$、$\mathcal{L}_{aggre}$、$\mathcal{L}_{neg}$进行梯度优化。在预测阶段，模型计算样本特征$x_i$与关系描述特征集合$y = \{y_1, \ldots, y_i, \ldots, y_k\}$中每个关系描述之间的特征相似分数，组成集合$Score_i = \{Score_{i1}(x_i, y_1), \ldots, Score_{ii}(x_i, y_i), \ldots, Score_{ik}(x_i, y_k)\}$。之后，从集合$Score_i$中找到最高分数$Score_{imax}$，将$y_{max}$对应的关系标签$r_{max}$作为样本语句$t_i$的预测结果，并与其头尾实体$t_i^{e1}$、$t_i^{e2}$组成关系三元组$(t_i^{e1}, t_i^{e2}, r_{max})$。

## 4 实验

### 4.1 实验设置

本文分别使用两个数据集对UFSC模型进行评估：FewRel (Han et al., 2018)和Wiki-ZSL (Chen and Li, 2021)。FewRel是一个基于Wikipedia的开放数据集，由许多数据标注者手动标注而成，其包含80种关系标签，每种关系有700个样本。目前FewRel广泛用于小样本学习任务。本文使其在训练阶段与预测阶段的关系标签集合不相交，将其应用于零样本学习任务。Wiki-ZSL是一个基于Wikidata知识库远程监督生成的数据集，其包含113种关系标签和93383个样本。与FewRel相比，Wiki-ZSL有着更多样的关系标签，但是其样本内部存在更多噪声。

与以往的研究者一样，本文选择5个关系标签作为验证集，$unseen \in \{5, 10, 15\}$为测试集中不可见关系标签的数量，并将剩余的关系标签作为可见关系标签对模型进行训练。同时，为了保证实验结果的可靠性，每组实验均重复进行5次，每次对关系标签重新选择，将平均结果作为最终实验结果。

### 4.2 基线模型

为了更好评估方法的有效性，本文分别与较为主流的生成式方法和最先进的基于匹配的零样本关系抽取方法进行对比。同时，还比较了最近较为流行的基于大模型的零样本关系抽取方法。

- ZS-BERT (Chen and Li, 2021)：一种基于匹配的零样本关系抽取方法，分别采用BERT和Sentence-Bert作为编码器对样本和关系描述进行编码。除了分类损失外，还使用对比学习的方法来优化特征空间表示。

- PromptMatch (Sainz et al., 2021)：一种SOTA的生成式零样本关系抽取方法，对样本进行细粒度建模。


(c) 2024 中国中文信息学会计算语言学专业委员会

- REPrompt (Chia et al., 2022)：一种基于生成关系伪数据的零样本关系抽取方法。
- SUMASK (Li et al., 2023)：一种基于大模型的零样本关系抽取方法，将关系抽取任务转换为多轮问答任务。通过设置不同的问题模板，利用大模型进行样本信息总结、二分类关系判断和语句匹配，最终对比语句匹配度完成对关系标签的抽取。
- RE-Mathing (Zhao et al., 2023)：一种SOTA的基于匹配的零样本关系抽取方法，提出一种细粒度匹配机制，充分利用实体中的特征信息。

### 4.3 实验细节

基线模型实验中大部分使用Bert作为样本编码器，为了公平比较并统一样本与描述的编码粒度，本文使用Bert-base-uncased作为样本语句和关系描述的编码器。选取*AdamW* (Loshchilov and Hutter, 2017)作为优化器，batchsize为24。$unseen \in \{5, 10, 15\}$时，FewRel数据集上epoch分别为2、3、3，Wiki-ZSL数据集上epoch分别为2、2、3。训练过程中，学习率为2e-5。同时，采用预热学习率策略，预热系数为0.2。超参数$\alpha$与$\beta$均为0.2，负样本随机采样数量$ra$为7，负样本检索数量$re$为5，最高聚合限度$margin$为0.95，最小相邻间隔$interval$为0.06。在$unseen \in \{5, 10\}$时，使用提示加关系定义组成关系描述；在$unseen = 15$时，仅使用关系定义作为关系描述。所有实验均使用NVIDIA GeForce RTX 3090进行。

### 4.4 主要实验结果

| *unseen* | 方法 | Wiki-ZSL | | | FewRel | | |
|---|---|---|---|---|---|---|---|
| | | *Prec.* | *Rec.* | $F_1$ | *Prec.* | *Rec.* | $F_1$ |
| **5** | ZS-Bert (Chen and Li, 2021) | 71.54 | 72.39 | 71.96 | 76.96 | 78.86 | 77.90 |
| | PromptMatch (Sainz et al., 2021) | 77.39 | 75.90 | 76.63 | 91.14 | 90.86 | 91.00 |
| | REPrompt (Chia et al., 2022) | 70.66 | 83.75 | 76.63 | 90.15 | 88.50 | 89.30 |
| | SUMASK (Li et al., 2023) | 75.64 | 70.96 | 73.23 | 78.27 | 72.55 | 75.30 |
| | RE-Matching (Zhao et al., 2023) | 78.19 | 78.41 | 78.30 | 92.82 | 92.34 | 92.58 |
| | **UFSC(ours)** | 85.84 | 84.43 | **85.11** | 92.74 | 92.31 | 92.53 |
| **10** | ZS-Bert (Chen and Li, 2021) | 60.51 | 60.98 | 60.74 | 56.92 | 57.59 | 57.25 |
| | PromptMatch (Sainz et al., 2021) | 71.86 | 71.14 | 71.50 | 83.05 | 82.55 | 82.80 |
| | REPrompt (Chia et al., 2022) | 68.51 | 74.76 | 71.50 | 80.33 | 79.62 | 79.96 |
| | SUMASK (Li et al., 2023) | 62.31 | 61.08 | 61.69 | 64.77 | 60.94 | 62.80 |
| | RE-Matching (Zhao et al., 2023) | 74.39 | 73.54 | 73.96 | 83.21 | 82.64 | 82.93 |
| | **UFSC(ours)** | 75.37 | 76.58 | **75.95** | 85.82 | 85.97 | **85.89** |
| **15** | ZS-Bert (Chen and Li, 2021) | 34.12 | 34.38 | 34.25 | 35.54 | 38.19 | 36.82 |
| | PromptMatch (Sainz et al., 2021) | 62.13 | 61.76 | 61.95 | 72.83 | 72.10 | 72.46 |
| | REPrompt (Chia et al., 2022) | 63.69 | 67.93 | 65.74 | 74.33 | 72.51 | 73.40 |
| | SUMASK (Li et al., 2023) | 43.55 | 40.27 | 41.85 | 44.76 | 41.13 | 42.87 |
| | RE-Matching (Zhao et al., 2023) | 67.31 | 67.33 | 67.32 | 73.80 | 73.52 | 73.66 |
| | **UFSC(ours)** | 68.61 | 68.86 | **68.72** | 76.02 | 75.14 | **75.57** |

Table 3: 模型在两个零样本关系抽取数据集上的主要实验结果。

UFSC在Wiki-ZSL和FewRel数据集的实验结果如Table 3所示。从实验结果中可以发现，在针对不同数量的不可关系*unseen*时，UFSC在FewRel与Wiki-ZSL上的表现均优于现有基线。与之前最佳的结果相比，UFSC在FewRel上$F_1$值平均提高1.6%，在Wiki-ZSL上$F_1$值平均提高3.4%。ZS-Bert是针对零样本关系抽取任务设计的，但是关系描述编码模块在训练过程中固定不变，对描述的特征表示较为单一，导致其性能不佳。PromptMatch使用自注意力机制对样本特征进行建模，UFSC并未对样本特征进行去噪或细粒度建模操作，但性能仍有较大提升。REPrompt在训练过程中使用新的关系伪数据对模型进行训练。与之相比，UFSC在训练过程中对数据的依赖性更小，且能够达到更优的性能。与SOTA方法RE-Matching相


(c) 2024 中国中文信息学会计算语言学专业委员会

比，UFSC统一样本语句与关系描述的特征空间，使其在预测阶段取得更佳的匹配性能。同时，UFSC的特征均匀化模块也可以带来更加均匀的特征空间分布，对相似关系标签的分辨能力更佳。SUMASK在抽取过程中引入大模型交互问答，并与UFSC一样使用语句匹配的方式完成关系抽取任务。然而UFSC并未使用大模型的外部知识库对特征信息进行总结，也并未使用大模型识别语句间的匹配程度，但是其性能较SUMASK有很大提升。这得益于统一特征空间的构建与特征均匀化操作，使得样本特征与描述特征相统一并都有着更为准确的特征表示。

### 4.5 消融实验

#### 4.5.1 关系描述重构策略消融实验

本节设置三组实验来讨论不同结构的关系描述对于模型性能的影响。Prompt组使用提示与关系定义组成关系描述,例如对于关系标签“Mother”，使用描述“the relation between head entity and tail entity is mother, mother is female parent of the subject”；Label组使用标签与关系定义组成关系描述，例如“mother is female parent of the subject”；NoPrompt组仅使用关系定义作为关系描述，如“mother: mother is female parent of the subject”。三组实验中，模型仅采用不同关系描述组成方法，其余部分均保持一致。实验结果如表Table 4所示。

| *unseen* | 方法 | Wiki-ZSL | | | FewRel | | |
|---|---|---|---|---|---|---|---|
| | | *Prec.* | *Rec.* | $F_1$ | *Prec.* | *Rec.* | $F_1$ |
| **5** | UFSC-Prompt | 85.84 | 84.43 | 85.11 | 92.74 | 92.31 | 92.53 |
| | UFSC-NoPrompt | 83.54 | 82.39 | 82.94 | 91.85 | 91.35 | 91.60 |
| | UFSC-Label | 83.60 | 81.90 | 82.71 | 91.30 | 90.67 | 90.99 |
| **10** | UFSC-Prompt | 75.37 | 76.58 | 75.95 | 85.82 | 85.97 | 85.89 |
| | UFSC-NoPrompt | 73.59 | 74.39 | 73.97 | 85.64 | 84.32 | 84.96 |
| | UFSC-Label | 73.58 | 74.48 | 74.01 | 85.72 | 84.80 | 85.25 |
| **15** | UFSC-Prompt | 67.44 | 68.94 | 68.16 | 74.92 | 74.70 | 74.79 |
| | UFSC-NoPrompt | 68.61 | 68.86 | 68.72 | 76.02 | 75.14 | 75.57 |
| | UFSC-Label | 66.62 | 67.52 | 67.03 | 74.92 | 74.33 | 74.62 |

Table 4: 关系描述不同构造方法的比较。

以数据集FewRel中的结果为例，在$unseen \in \{5, 10\}$时，Prompt组相比于NoPrompt组，精准率、召回率和$F_1$值均有所提升。这得益于描述中所含特征信息更为多样。而在不可见关系较多时，NoPrompt组由于描述中的噪声较少，泛化能力优于Prompt组。因此，关系描述策略采用在预测空间较小时，重构关系描述以提升描述信息的丰富性；同时在预测关系较多时，仅采用关系定义作为关系描述，以避免不必要噪声的引入。

对比Prompt组与Label组可以发现，无论*unseen*取何值，Prompt组的实验结果均优于Label组。相比于仅强调标签信息，引入关系作用信息可以帮助模型在低资源的场景下获得更多粒度的可学习信息。证明所引入提示的优越性。

| 关系标签 | 方法 | Wiki-ZSL | | | FewRel | | |
|---|---|---|---|---|---|---|---|
| | | *Prec.* | *Rec.* | $F_1$ | *Prec.* | *Rec.* | $F_1$ |
| Country | UFSC-Prompt | 77.32 | 65.17 | 70.72 | 94.92 | 34.71 | 50.84 |
| | UFSC-NoPrompt | 73.30 | 25.83 | 38.19 | 97.59 | 23.14 | 37.41 |
| Country of citizens | UFSC-Prompt | 71.23 | 98.50 | 82.67 | 63.78 | 95.86 | 76.60 |
| | UFSC-NoPrompt | 57.37 | 95.89 | 71.79 | 60.14 | 96.57 | 74.12 |

Table 5: $unseen = 10$时，关系描述重构对于相似关系标签的比较。

并且，引入关系作用信息可以帮助模型更准确地识别相似关系标签。如表Table 5所示，在$unseen = 10$时，Prompt组对于关系标签“Country”的召回率相较NoPrompt组有着大幅提升，即Prompt组对于“Country”标签的正样本识别率更高。这得益于引入关系作用信息后模型对于“Country”标签的关系描述特征有着更准确地表示，进而提高“Country”标签的关系描述与


(c) 2024 中国中文信息学会计算语言学专业委员会

其正样本在特征空间上的同义程度，使得模型对于“Country”与“Country of citizens”这对相似标签的有着更强的区分能力。

#### 4.5.2 特征均匀化模块消融实验

为探究描述特征均匀化方法与负样本检索方法对模型性能的影响，本节设置三组消融实验。实验结果如Table 6所示。NoHomogenization组完全去除特征均匀化模块；OnlyDistinguish组仅去除负样本检索方法；OnlyRetriever组仅去除描述均匀化方法。

| *unseen* | 方法 | Wiki-ZSL | | | FewRel | | |
|---|---|---|---|---|---|---|---|
| | | *Prec.* | *Rec.* | $F_1$ | *Prec.* | *Rec.* | $F_1$ |
| **5** | UFSC | 85.84 | 84.43 | 85.11 | 92.74 | 92.31 | 92.53 |
| | NoHomogenization | 83.74 | 81.47 | 82.55 | 89.76 | 88.98 | 89.37 |
| | OnlyDistinguish | 84.45 | 83.33 | 83.87 | 90.92 | 89.99 | 90.45 |
| | OnlyRetriever | 84.25 | 82.15 | 83.15 | 91.54 | 91.06 | 91.30 |
| **10** | UFSC | 75.37 | 76.58 | 75.95 | 85.82 | 85.97 | 85.89 |
| | NoHomogenization | 73.16 | 74.53 | 73.82 | 85.46 | 84.85 | 85.15 |
| | OnlyDistinguish | 74.70 | 75.68 | 75.18 | 84.78 | 84.34 | 84.56 |
| | OnlyRetriever | 73.72 | 74.26 | 73.98 | 86.04 | 85.44 | 85.73 |
| **15** | UFSC | 68.61 | 68.86 | 68.72 | 76.02 | 75.14 | 75.57 |
| | NoHomogenization | 64.91 | 64.72 | 64.79 | 76.11 | 74.55 | 75.31 |
| | OnlyDistinguish | 68.09 | 68.28 | 68.16 | 73.97 | 73.07 | 73.51 |
| | OnlyRetriever | 67.33 | 67.10 | 67.19 | 75.17 | 74.55 | 74.85 |

Table 6: 特征均匀化模块不同构造方法的比较。

实验结果表明，在大部分情况，使用描述特征均匀化方法与负样本检索均可以提升模型预测性能。并且，在一些场景下，两种方法相互补充，同时使用可以更加显著的提升模型性能。而在Wiki-ZSL数据集中，描述均匀化方法对模型性能有着更显著的提高。这是由于Wiki-ZSL中的样本存在更多噪声。因此，对描述特征的分辨可以将关系标签相同的样本特征进行聚类，不同标签的样本特征在空间上的分布间距更大，减小样本噪声对匹配过程的影响。

#### 4.5.3 统一特征空间构建消融实验

| *unseen* | 方法 | Wiki-ZSL | | | FewRel | | |
|---|---|---|---|---|---|---|---|
| | | *Prec.* | *Rec.* | $F_1$ | *Prec.* | *Rec.* | $F_1$ |
| **5** | UFSC | 85.84 | 84.43 | 85.11 | 92.74 | 92.31 | 92.53 |
| | UFSC-LabelLoss | 73.85 | 70.48 | 72.10 | 78.66 | 76.37 | 77.48 |
| | 1-Encoder | 64.72 | 60.34 | 62.34 | 73.88 | 60.60 | 66.10 |
| | 2-Encoder | 57.44 | 55.32 | 56.33 | 83.45 | 81.16 | 82.26 |
| **10** | UFSC | 75.37 | 76.58 | 75.95 | 85.82 | 85.97 | 85.89 |
| | UFSC-LabelLoss | 51.01 | 51.97 | 51.47 | 63.45 | 61.80 | 62.56 |
| | 1-Encoder | 50.72 | 42.78 | 46.13 | 64.01 | 46.12 | 53.56 |
| | 2-Encoder | 63.40 | 61.31 | 62.24 | 71.25 | 65.94 | 68.45 |
| **15** | UFSC | 68.61 | 68.86 | 68.72 | 76.02 | 75.14 | 75.57 |
| | UFSC-LabelLoss | 51.98 | 50.92 | 51.43 | 56.09 | 54.19 | 55.08 |
| | 1-Encoder | 51.41 | 41.05 | 45.45 | 53.10 | 41.13 | 46.24 |
| | 2-Encoder | 55.04 | 52.46 | 54.65 | 61.44 | 56.58 | 58.82 |

Table 7: 是否统一特征空间方法的比较。

本节研究特征空间是否统一对模型性能的影响。消融实验的结果如Table 7所示。1-Encoder组使用可学习的Bert对样本进行编码，使用固定的Sentence-Bert (Reimers and Gurevych, 2019)对描述进行编码；2-Encoder组使用两个独立的Bert分别对样本与描述进行


(c) 2024 中国中文信息学会计算语言学专业委员会

编码。带有LabelLoss的实验组，使用关系标签多分类损失替代特征相似损失，以验证特征相似损失为模型性能带来的提升。

在所有场景下，UFSC的性能均优于1-Encoder组。这是由于统一的特征空间帮助相同标签的样本与描述获得较为统一的特征表示，使得二者在后续的特征相似性计算中取得较好的成绩。对比1-Encoder组和2-Encoder组结果可以发现，可学习描述编码模块的性能在大部分场景下优于固定的描述编码模块，可以更为充分地学习描述中的特征信息。通过对比UFSC和LabelLoss组结果可以发现，相较于关系标签多分类损失，特征相似损失提升标签相同的样本特征与描述特征在特征空间上的相似性，在统一特征空间的场景下对模型性能有着较大提升。

## 5 结论

本文提出了一种基于统一特征空间构建的零样本关系抽取方法。通过统一样本语句与关系描述的编码模块并引入特征相似度损失，模型构建了样本与描述相统一的特征空间。同时，设置特征均匀化模块并引入负样本检索，帮助模型实现特征分布更加均匀的特征空间构建。

与当今的生成式大模型相比，UFSC体量较小可以在本地部署，便于在现有基础上进行微调，可以较好应用于某一特定领域。但是，大模型作为一种强大的外部知识库，可以在模型零样本学习的过程中，为模型补充更多的外部知识。UFSC中并未使用任何外部知识，因此其在零样本关系抽取性能仍具备提升空间。

而更丰富的特征信息可能带来更多的噪声，如何进行特征去噪仍然是一个值得研究的方向。大模型所提供的外部知识或许可以帮助识别特征中的噪声，进一步提高UFSC在抽取过程中的特征简洁度。同时，零样本关系抽取要求模型具备较好的泛化能力。因此，特征去噪是否同样具备可迁移性也是一个值得探讨的问题，将此留作未来的工作。

## 参考文献

Chih-Yao Chen and Cheng-Te Li. 2021. Zs-bert: Towards zero-shot relation extraction with attribute representation learning. *arXiv preprint arXiv:2104.04697*.

Yew Ken Chia, Lidong Bing, Soujanya Poria, and Luo Si. 2022. Relationprompt: Leveraging prompts to generate synthetic data for zero-shot relation triplet extraction. *arXiv preprint arXiv:2203.09101*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Zhijiang Guo, Guoshun Nan, Wei Lu, and Shay B Cohen. 2021. Learning latent forests for medical relation extraction. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3651–3657.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. Fewrel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. *arXiv preprint arXiv:1810.10147*.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. *arXiv preprint arXiv:1706.04115*.

Guozheng Li, Peng Wang, and Wenjun Ke. 2023. Revisiting large language models as zero-shot relation extractors. *arXiv preprint arXiv:2310.05028*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Abiola Obamuyide and Andreas Vlachos. 2018. Zero-shot relation classification as textual entailment. In *Proceedings of the first workshop on fact extraction and VERification (FEVER)*, pages 72–78.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Oscar Sainz, Oier Lopez de Lacalle, Gorka Labaka, Ander Barrena, and Eneko Agirre. 2021. Label verbalization and entailment for effective zero-and few-shot relation extraction. *arXiv preprint arXiv:2109.03659*.


(c) 2024 中国中文信息学会计算语言学专业委员会

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. *arXiv preprint arXiv:1906.03158*.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Conference on empirical methods in natural language processing*.

Jun Zhao, Wenyu Zhan, Xin Zhao, Qi Zhang, Tao Gui, Zhongyu Wei, Junzhe Wang, Minlong Peng, and Mingming Sun. 2023. Re-matching: A fine-grained semantic matching method for zero-shot relation extraction. *arXiv preprint arXiv:2306.04954*.

CCL 2024

# 多机制整合的中文医疗命名实体识别

**王珊珊[1],张焜元[2,3,4],闫蓉[2,3,4*]**

[1]延安大学经济与管理学院 陕西, 延安, 716000

[2]内蒙古大学计算机学院 内蒙古, 呼和浩特, 010021

[3]蒙古文智能信息处理技术国家地方联合工程研究中心 内蒙古, 呼和浩特, 010021

[4]内蒙古自治区蒙古文信息处理技术重点实验室 内蒙古, 呼和浩特, 010021

wangshanshan@126.com, 11467705@qq.com, csyanr@imu.edu.cn

## 摘要

在互联网在线医疗领域，由于大多数患者缺乏医学培训，以及不同学科病理特征的复杂性，医患对话文本中的医学命名实体呈现出长且多词的句法特点，给命名实体识别算法提出了新的挑战。为解决这一问题，本研究融合多个不同粒度的扩张卷积机制，构建了Flat-Lattice-CNN模型。该模型不仅考虑字符和词语的语义信息以及它们的绝对和相对位置信息，还提取跨越不同距离的多个字符/词语的共现依存关系特征，以此提高医学长命名实体的识别精度。实验结果表明，本文提出的模型在所评估数据集的命名实体识别任务上有普遍性的性能提升，尤其是在以长实体为主的中文医疗数据集CTDD上，该模型的$F1$值提升了约2%，具有更优的表现。

**关键词：** 长且多词医学命名实体；中文医疗命名实体识别；扩张卷积

# Infusing multi-schemes for Chinese Medical Named Entity Recognition

**Shanshan Wang[1],Kunyuan Zhang[2,3,4],Rong Yan[2,3,4*]**

[1]Economics and Management School, Yan'an University, Yanan, 716000, China

[2]College of Computer Science, Inner Mongolia University, Hohhot, 010021, China

[3]National and Local Joint Engineering Research Center of Intelligent Information Processing Technology for Mongolian, Hohhot, 010021, China

[4] Inner Mongolia Key Laboratory of Mongolian Information Processing Technology, Hohhot, 010021, China

wangshanshan@126.com, 11467705@qq.com, csyanr@imu.edu.cn

## Abstract

In the field of internet-based healthcare, the complexity of pathology features across various disciplines, coupled with the lack of medical training among most patients, results in medical named entities in doctor-patient dialogue texts exhibiting long and

第二十三届中国计算语言学大会论文集，第382页-第393页，太原，中国，2024年7月25日至28日。
卷1：主会论文
(c) 2024 中国中文信息学会计算语言学专业委员会

multi-word syntactic patterns, posing new challenges to named entity recognition algorithms. To address this issue, this study integrates multiple dilation convolution mechanisms of different granularities to construct the Flat-Lattice-CNN model. This model not only considers the semantic information of characters and words, as well as their absolute and relative positional information but also extracts relationship features between characters/words spanning different distances to improve the recognition accuracy of long medical named entities. Experimental results show a significant performance boost in the task of recognizing medical named entities on all evaluation dataset, especially on CTDD with a 2% increase in $F1$ score.

**Keywords:** Long multiword named entities , Chinese Medical Named Entity Recognition , Dilated Convolution

## 1 引言

命名实体识别(Named Entity Recognition, NER)是自然语言处理的基础任务之一，其主要功能是在文本中识别特定概念的实体，对于机器翻译、信息提取、问答系统等领域至关重要。医学命名实体识别旨在从海量的非结构化医疗数据中提取关键信息，如药物、手术、症状、身体、疾病等，为医学研究发展和智慧医疗系统普及提供基础支持(张帆 and 王敏, 2017)。近年来，随着信息技术的不断进步和人们服务意识的提高，互联网在线医疗正逐渐成为传统面对面医疗的重要补充(罗歆然 et al., 2024)。在互联网在线医疗平台上，诊疗期间的医患交流对话多以文本形式记录并保存。但由于大多数患者缺乏医学培训，以及医学不同学科病理特征的复杂性，他们往往在提供病情和病史时采用较长的表达方式。因此，相关的命名实体表现出长且多词的特征，例如，妇产科领域中“月经过后老是再沥沥拉拉一周左右褐色的分泌物”是一个长且多词的医学症状实体。相比较下，已有研究中的医学命名实体通常以较短和较少词语标记呈现。举例而言，“发烧”被视为一个具有完整且独立医学意义的症状实体，而“断断续续地发烧”也是一个具有完整且独立医学意义的症状实体。尽管它们可能对应着不同的病理情况，然而已有研究却通常将它们等同对待，即“断断续续地发烧”内容中，“断断续续”这一部分词语多数情况下会被忽略标记和识别。

目前，医疗文本命名实体识别的研究主要集中在电子病历、医学文献、医学书籍等领域，而对互联网在线医疗问答文本的研究则较为有限(Li et al., 2020)。虽然已有方法例如BiLSTM-CRF(Luo et al., 2018)、Bert-CRF(Souza et al., 2019) 和LeBert-CRF(Liu et al., 2021)等，在

©2024 中国计算语言学大会 根据《Creative Commons Attribution 4.0 International License》许可出版
**基金项目：**内蒙古自然科学基金(2023MS06023)

医疗命名实体识别领域获得了有效结果，但其应用到互联网在线医疗问答中文文本领域时存在一定局限。具体原因如下：首先，它们主要是针对较短的命名实体识别任务开发的，但在互联网在线医疗医患对话文本中，相关医学命名实体表现出长且多词的句法特点；其次，大多数模型聚焦在英文文本中，而英文和中文的句法特征有较大区别，前者基于字符特征，后者除了字符对词语特征也有较强依赖；最后，虽然命名实体识别普遍被形式化为序列标记问题，但其本质上是对实体内部词之间关系的建模(Li et al., 2022)，然而迄今为止，算法(例如，注意力机制)(Vaswani et al., 2017)更多捕捉的是标记(字符或词语)之间的两两关系信息，而长且多词命名实体中的多词共现依存关系特征仍然缺乏有效的捕获机制。

为解决上述问题，本研究针对互联网医患对话文本，提出了一种整合Flat-Lattice和多粒度扩张卷积网络机制的新模型：Flat-Lattice-CNN。在这个模型中，Flat-Lattice结构简化了模型的复杂性，将其结构设计成了平面形式，不仅降低了模型的复杂度，还充分利用了字符和词级别的信息。虽然Flat-lattice机制可以提取实体中的标记(字符或词语)对之间的长依赖关系，但其主要用于捕获标记之间两两的关系信息。因此，针对长且多词的长医学实体，为了建模实体内多词长距离的共现依存关系特征，本研究整合多粒度扩张卷积机制。实验结果表明，该模型在数据集的命名实体识别任务上有普遍性的性能提升。

## 2 相关工作

### 2.1 命名实体识别

命名实体识别（NER）长期以来一直是自然语言处理（NLP）领域的基础任务之一，有着广泛的应用，包括但不限于机器翻译、信息提取、问答系统等。从命名实体识别问题的分析角度，近年来，NER的研究逐渐从最初的平铺式NER(Flat NER)，演变到了嵌套NER(Nested NER)和不连续NER(Discontinuous NER)(Li et al., 2022; Zheng et al., 2019)。

相较于通用领域的命名实体识别，医学命名实体识别任务需要结合医学领域的语言特点和医学语义知识。经过文献梳理，我们发现多数已有医学命名实体工作主要聚焦在输入表示方面，例如Li等采用Lattice-LSTM整合词汇信息，并基于ELMo学习电子病历中的上下文信息(Zhang and Yang, 2018)；(Luo et al., 2018)利用字符串匹配方法将疾病字典与字符进行配对，并提出了结合字典注意力层的BiLSTM-CRF模型；(罗歆然 et al., 2024)提出了一种BERT-BiLSTM-CRF模型，该模型融合了词汇和字根特征，利用BERT模型增强了中文临床记录文本的上下文语义信息。除此之外，也有关注医学文本中的嵌套实体识别的工作(Fei et al., 2021)。近年来，学者们注意到在医学文本中出现不连续命名实体的现象，例如“腿和胳膊疼”。针对这类问题，相关研究提出引入新的标记标签，例如“感觉(疼)”来处理这种不连续实体的情况，这类标签的数量随着语料库的变化，需要及时进行更新和丰富(Dai et al., 2020)。

鉴于先前研究和互联网医患对话文本特点，我们发现长且多词的医学命名实体识别问题尚

CCL 2024

未得到重视。例如“断断续续地发烧”中“断断续续”容易被已有医学命名实体识别工作所忽略。考虑到命名实体识别本质上是对实体标记(字符或词)依赖关系的建模，以及中文文本的特点，本研究整合Flat-Lattice和多粒度膨胀卷积机制，提出Flat-Lattice-CNN模型，其中多粒度扩张卷积可有效地捕获跨越短和长距离的多标记(字符或词语)共现依存的关系信息，用以识别长且多词医学命名实体。

### 2.2 Flat-Lattice

相比英文，中文的命名实体识别更加困难，因为它涉及到词语的分词问题。最近的研究已经证明了栅格结构在利用词语信息和避免分词错误传播方面具有优势。在栅格结构中，可以将一个句子与词典进行匹配，以获取其中的潜在词语，这些词语对于命名实体识别可能非常重要。具体而言，栅格中的每个节点都是标记(一个字符或一个潜在词语)，其中，词语并非按顺序排列，而是由词语的第一个字符和最后一个字符决定其位置。基于栅格的模型使用注意力机制在每个位置融合可变数量的词语节点。然而，由于这种栅格结构的动态性，相关方法不能充分利用GPU的并行计算。因此，李等人提出了Flat-Lattice，该模型基于Transformer将栅格结构转换为平面格(flat-lattice)，通过设计的头(head)和尾(tail)标记的位置编码，该模型可无损复原栅格结构。因此，可以直接使用Transformer来完全模拟栅格输入。不仅如此，Transformer的自注意力和全连接机制使字符可以直接与任何潜在词进行交互，并模拟序列中标记之间的不同距离依赖关系(Li et al., 2020)。

### 2.3 CNN

卷积神经网络(Convolutional Neural Networks,CNN)(Zeng et al., 2014)是一种结构特殊的人工神经网络，主要应用在视频分类、图像识别、图像标注、人脸检测、信息抽取、文本处理和语音识别等方面，在实际应用中达到很好的效果。卷积神经网络一般由输入层、卷积层、池化层、激活层以及顶端的全连接层、损失函数层组成。卷积层的目的是提取不同特征，位置较低的卷积层提取低级特征，位置较高的卷积层提取高级特征。其中，卷积层的卷积核感受视野范围由膨胀率控制，即膨胀率越高，卷积核的感受视野越大，如果是在文本处理领域，则能建模跨越长距离的多词共现依存关系特征。

## 3 Flat-Lattice-CNN模型

本文提出的Flat-Lattice-CNN模型主要由三部分组成，分别是输入层、Self-Attention+FFN层和标签预测层，整体结构如图1所示。首先，利用Flat-Lattice Transformer获取输入句子的标记(字符或词语)嵌入向量和标记的绝对、相对位置向量，进一步在此基础上采用不同膨胀率的卷积核将跨越不同距离的多词共现依存关系的特征转化成向量，然后融合这三类型向量，得到最终的输入表示。其次,将输入层得到的向量输入至Self-Attention+FFN层，层


(c) 2024 中国中文信息学会计算语言学专业委员会

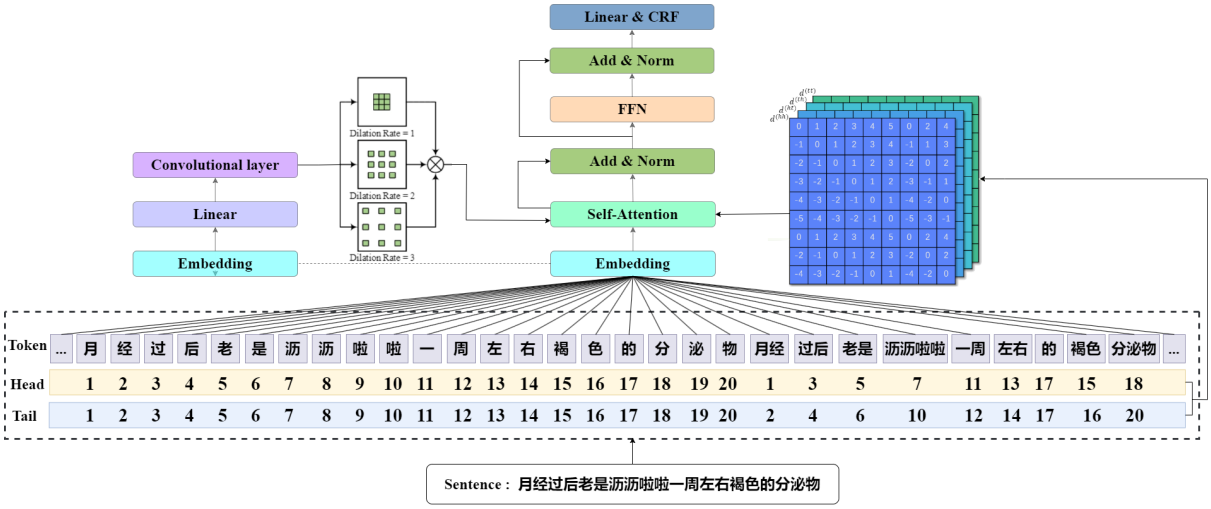



图 1: FLAT-Lattice-CNN模型的整体架构图

中使用Self-Attention和FFN学习标记(字符或词语)的上下文环境，然后使用残差结构得到最终的语义特征。最后,将最终的语义特征输入到以CRF为函数的标签预测层中。接下来将详细说明各个部分。

### 3.1 Flat-Lattice Transformer

Flat-Lattice Transfomer先将字符与词汇表生成的格子展开为扁平化span序列，之后采用Transformer融合位置信息得出每个标记的嵌入向量。具体的，每一个span对应一个标记(字符或词语)、一个头(head)和一个尾(tail)的位置信息，就像图1虚线框中所示。标记是一个字符或单词。头和尾指示了标记在原序列中第一个和最后一个字符的位置索引，也就是它们表示了标记在格子中的位置。Transformer采用全连接的自注意力机制和前馈网络来建模序列中的标记(字符或词语)之间的依赖关系。

为了保留位置信息,Flat-Lattice Transformer为序列中的每个标记引入了位置表示，形成对应的每个span。具体而言，为每一个标记（字符或单词）分配了两个位置索引：头位置和尾位置，通过这些位置索引，可以从一个标记序列中重构出一个格子。输入序列中的任意两个SPAN$x_i$和$x_j$，它们之间存在相交、包含和分离三种关系，由span的头和尾共同决定。因此，这种位置机制可以同时表示两个标记绝对和相对位置信息，任意两个SPAN$x_i$和$x_j$的相对位置可

以用公式1计算：

$$\begin{aligned} d_{ij}^{(hh)} &= head\left[i\right] - head\left[j\right] \\ d_{ij}^{(ht)} &= head\left[i\right] - tail\left[j\right] \\ d_{ij}^{(th)} &= tail\left[i\right] - head\left[j\right] \\ d_{ij}^{(tt)} &= tail\left[i\right] - tail\left[j\right] \end{aligned} \tag{1}$$

其中，$d$表示头部之间的距离$i$和$j$具有类似的含义。

SPAN的相对位置编码是四个距离的简单非线性变换，计算方法如2所示：

$$R_{ij} = \mathrm{ReLU}\left(W_r\left(P_{d_{ij}^{(hh)}} \oplus P_{d_{ij}^{(th)}} \oplus P_{d_{ij}^{(ht)}} \oplus P_{d_{ij}^{(tt)}}\right)\right) \tag{2}$$

其中$Wr$是可学习参数，$\bigoplus$表示串联运算符，$pd$的计算方法如Vaswani等人所述(Vaswani et al., 2017)，可通过公式3来计算：

$$\begin{aligned} P_d^{(2k)} &= \sin\left(d/10000^{2k/d_{model}}\right) \\ P_d^{(2k+1)} &= \cos\left(d/10000^{2k/d_{model}}\right) \end{aligned} \tag{3}$$

其中，$d$和$k$表示位置编码的维度索引。然后，使用自注意力的变体来利用相对SPAN的位置编码，以进一步提高模型对序列信息的建模能力。通过融合的标记(字符或词语)和它们的相对、绝对位置信息，并经过transformer机制后，得到了输入序列中每个span的嵌入向量$w_F$。

### 3.2 Dilated Convolution CNN

在前述嵌入向量$w_F$基础上，本研究引入了不同膨胀率的卷积机制，以此获取实体内跨越不同距离的多词共现依存关系特征向量$w_c$，然后与$w_f$拼接成为最终的输入嵌入向量$[w_f; w_c]$。如公式4所示，卷积机制具有局部感受视野和平移不变性的机制，局部感受野是指机制能够专注于局部特征的提取，而不需要考虑整个输入的复杂性；平移不变性指的是是指特征的识别不受其绝对位置的影响。 Dilated Convolution CNN结构包括以下几个关键组件：

1.基础卷积层：首先通过一个具有可选丢弃率的一维卷积层来对输入进行特征提取。这有助于将输入数据映射到一个更高维度的表示空间，并通过RELU激活函数进行非线性转换。这一步旨在为后续的多粒度扩张卷积层提供一个丰富的输入表示。

2.多粒度扩张卷积(Li et al., 2022)：为了捕获多词的共现依存关系信息，本文引入了多个具有不同扩张率的卷积核。每个卷积核都具有自己的感受野大小，使其能够有效地捕获不同范围内的语义信息。具体地，假设以一个变量$a$来衡量空洞卷积的扩张系数，则加入空洞之后的实际卷积核尺寸与原始卷积核尺寸之间的关系为公式4：

$$K = k + \left(k - 1\right)\left(a - 1\right) \tag{4}$$


(c) 2024 中国中文信息学会计算语言学专业委员会

其中，$k$为原始卷积核大小，$a$为卷积扩张率(dilation rate)；$K$为经过扩展后实际卷积核大小。卷积后得到的尺寸用公式5计算：

$$W = \frac{w-k+2p}{s}+1 = \frac{w-k-(k-1)(a-1)+2p}{s}+1 \tag{5}$$

其中，$w$为输入尺寸，$k$为原始卷积核大小，$a$为卷积扩张率，$s$ 为stride，$p$是padding。

3.残差连接(He et al., 2016)：为了确保信息的流动和保留原始输入的重要信息，本文在多粒度扩张卷积之后采用了残差连接的策略。

### 3.3 Self-Attention+FFN

本研究在整合Flat-Lattice Transformer和多通胀率的卷积机制基础上获取了输入序列的嵌入表示，之后将嵌入表示的序列输入到自注意力(Self-Attention)和全连接层(FFN)，使得模型在输入层基础上进一步提升处理每个元素考虑整个序列上下文信息的能力，从而更好地理解元素与序列中其他元素之间的关系。

### 3.4 预测层

标签预测层采用的是条件随机场模型 (Condi-tional Random Field,CRF)(Lafferty et al., 2001)。CRF是一种判别式的无向图模型，通过研究标签之间的关系，获得全局最优的标签序列。

## 4 实验结果与分析

### 4.1 实验数据

为验证本文模型在中文命名实体识别任务中的有效性，在IMCS-2021(Chen et al., 2023)和CTDD(Wang et al., 2023)两个公共中文医疗NER数据集进行实验，两个数据集的描述信息如表 1所示。CTDD是妇产科领域互联网医患对话中文数据集，IMCS-NER则是儿科领域的在线医患对话中文数据集。如表1所示，CTDD的实体平均长度为4.33，IMCS-2021则是2.62，因此可以得出CTDD的长实体比例比IMCS-2021更高。

IMCS-2021数据集涵盖症状、检查、药物名称等5种实体类型，本文从中随机抽取73,603条数据作为训练集，12,517条数据作为验证集，12,332条数据作为测试集。CTDD数据集包含时间、疾病和症状等6种实体类型，本文随机抽取21,633条作为训练集，3,757条作为验证集，4,847作为测试集。其中，CTDD数据集实体类型占比如图 2所示。

在本文命名实体识别实验评估任务中，参数设置如表 2所示。

CCL 2024

| | CTDD | IMCS-NER |
|---|---|---|
| 所有命名实体的数量 | 63560 | 74698 |
| 实体平均长度 | 4.33 | 2.62 |
| 总字符数 | 1,700,392 | 1,621,161 |
| 标记字符与总字符的比例 | 16.2% | 12.1% |
| 对话的平均字数 | 713.55 | 589.04 |

表 1: CTDD数据集与IMCS-NER数据集的描述信息

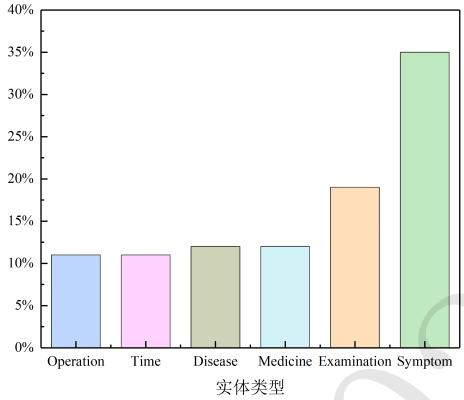



图 2: CTDD数据集实体类型分布图

### 4.2 评价指标

本文实验结果使用$F1$值作为评价指标，计算如公式6所示:

$$\begin{aligned} P &= \frac{\text{预测正确的实体集}}{\text{得到的实体总数}} \times 100\% \\ R &= \frac{\text{预测正确的实体集}}{\text{数据集实体数}} \times 100\% \\ F_1 &= \frac{2 \times P \times R}{P + R} \times 100\% \end{aligned} \tag{6}$$

### 4.3 基线模型

(1)BiLSTM-CRF(Luo et al., 2018)模型采用双向长短期记忆网络(Bidirectional Long Short-Term Memory，BiLSTM)来捕获输入序列中的上下文信息，并结合条件随机场(Conditional Random Field，CRF)来对序列标注问题进行建模。

(2)Bert-CRF(Souza et al., 2019)模型结合了预训练的 BERT 模型和条件随机场(constant ratefactor，CRF)，利用 BERT 模型强大的上下文理解能力，对输入序列进行表示学习，并通过 CRF 层进行标签序列的全局优化。

| 参数 | 数量 |
|---|---|
| 学习率 | 6e-4 |
| batch | 10 |
| epoch | 50 |
| 注意力头的个数 | 8 |
| 注意力头的维度 | 20 |
| 输入维度 | 160 |
| 通道数 | 160 |
| Dropout | 0.5 |

表 2: 参数设置

(3)LeBert-CRF(Liu et al., 2021)模型在 Bert-CRF 模型的基础上进行了改进，引入了领域专属的预训练方法。

(4)Lattice LSTM(Zhang and Yang, 2018)模型是一种基于格子结构的神经网络模型，它在模型的输入端引入了字符级别的信息，并通过格子结构的设计，使得模型能够更好地捕获单词内部字符之间的依赖关系。

### 4.4 实验结果分析

为更好观察本文模型对中文医疗命名实体识别的效果，分别用BiLSTM-CRF、Lattice LSTM、Bert-CRF、LeBert-CRF、FLAT-Lattice-Transformer模型作为对照组，分别在IMCS-2021和CTDD两个公共中文医疗NER数据集进行实验。对照组模型参数均为默认值，不同模型的中文命名实体识别效果如表3所示。

| | CTDD | | | IMCS-2021 | | | weibo | | |
|---|---|---|---|---|---|---|---|---|---|
| 模型 | P(%) | R(%) | F1(%) | P(%) | R(%) | F1(%) | P(%) | R(%) | F1(%) |
| Bert-CRF | 53.09 | 52.75 | 52.92 | 88.46 | 92.35 | 90.37 | - | - | - |
| LeBert-CRF | 52.63 | 54.25 | 53.42 | 86.53 | 92.91 | 89.60 | - | - | - |
| BiLSTM-CRF | 46.26 | 50.34 | 48.21 | 85.67 | 88.72 | 87.61 | - | - | - |
| Lattice LSTM | **57.14** | 55.17 | 56.14 | 89.37 | 90.84 | 90.10 | 53.04 | 62.25 | 58.79 |
| FLAT-Lattice Transformer | 53.62 | 56.47 | 55.14 | 88.83 | 93.43 | 91.07 | 69.84 | 77.37 | 73.41 |
| Flat-Lattice-CNN (ours) | 55.63 | **59.10** | **57.31** | **89.44** | **93.51** | **91.43** | **70.29** | **79.69** | **74.69** |

表 3: 在CTDD、IMCS、weibo数据集上的实验结果

虽然长且多词的实体是医学领域的文本特点，但是其它领域也可能有长实体的存在，因此本文对微博中文数据集在算法FLAT-Lattice Transformer和Flat-Lattice-CNN等进行了对比实验评估，如表3。由表3可以观察到，该模型有普遍性的性能提升，尤其时在长实体数据集上，

CCL 2024

$F1$值提升了2%，这也在一定程度上证明了本文模型的有效性。

为更好观察本文中多粒度扩张卷积的对中文医疗命名实体识别效果的影响，$\lambda$为膨胀率，即Dilation分数，本文分别建立了$\lambda$=[1]，$\lambda$=[1,2]，$\lambda$=[1,2,4]作为对照组模型，其他参数均为默认值，不同$dilation$($epoch$=15情况下的在CTDD数据集的命名实体识别效果如表4所示，由表中可以观察到膨胀率$\lambda$=[1,2,4]时$F1$分数和准确率($ACC$)值均有所提高。在模型中增加膨胀率4的卷积核，我们发现实验性能比低膨胀率的时候性能提升。

| Dilation($\lambda$) | $F1$(%) | $ACC$(%) |
|---|---|---|
| [1] | 53.40 | 86.53 |
| [1,2] | 54.50 | 87.61 |
| [1,2,4] | 55.35 | 87.75 |
| [1,2,4,8] | 52.77 | 87.70 |

表 4: 膨胀率的对比

| 句子 | 超过38.5度才给孩子用退烧药 |
|---|---|
| Bert-CRF | 超过38.5度(Symptom)才给孩子用退烧药(Drug) |
| Lattice LSTM | 超过38.5度(Symptom)才给孩子用退烧药(Drug) |
| Ours | 超过38.5度(Symptom)才给孩子用退烧药(Drug) |

表 5: IMCS-2021数据集命名实体识别效果实例

进一步地，本文列举出一个在以短实体为主的IMCS-2021数据集上的实例，如表5所示，其中Drug代表药物，退烧药是一个药物命名实体，Symptom代表症状。由表5可以观察到，Lattice LSTM模型将“38.5度”识别为短的单独的症状命名实体，而本文模型将“超过38.5度”成功识别为一个较长的症状命名实体，由表5可以直观地观察到本文模型对长实体识别改进的有效性。

由于本文采用了卷积层来捕获实体内多标记共现依存关系特征，在实验中我们观察到，模型捕获信息的速度更快，在$epoch$=15的情况下在CTDD数据集中$F1$分数的训练结果已经达到了52%（约为最终结果的91.2%）而FLAT-Lattice-Transformer的$F1$分数达到了0.48%（约为最终结果的87.2%），模型推理速度对照结果如图3所示。

## 5 结束语

本文介绍了一种基于 Flat-Lattice-CNN 的中文医疗命名实体识别方法，通过结合 Flat-Lattice 结构和不同粒度的膨胀卷积机制，有效地捕获了医疗文本中长且多词的命名实体中的多词共现依存关系信息。一定程度上克服了以注意力机制为基础的算法在这方面的局限。实验


(c) 2024 中国中文信息学会计算语言学专业委员会

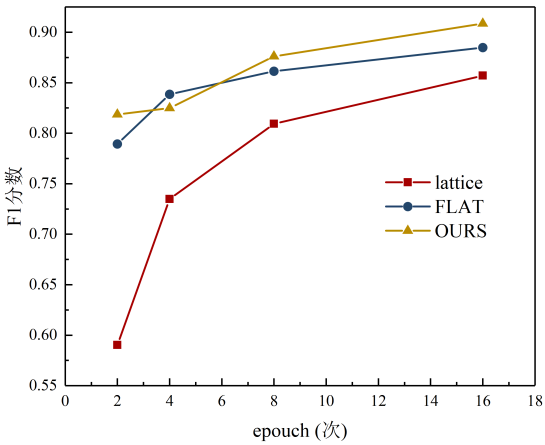



图 3: 模型推理速度对照结果图

结果表明，相比传统方法，本文模型在$F1$值和收敛速度上均有一定的提升，丰富了医疗信息处理领域的思路和方法。将来，我们会探索长且多词命名实体识别效果的语言学理论理解和效果的进一步优化，以应对更加复杂和多样化的医疗文本数据。

## 致谢

本项研究获得了国家自然科学基金(编号:72261032)、内蒙古自然科学基金(编号:2023MS06023)和陕西社会科学基金(编号:2022N005)资助。

## 参考文献

Wei Chen, Zhiwei Li, Hongyi Fang, Qianyuan Yao, Cheng Zhong, Jianye Hao, Qi Zhang, Xuanjing Huang, Jiajie Peng, and Zhongyu Wei. 2023. A benchmark for automatic medical consultation system: frameworks, tasks and datasets. *Bioinformatics*, 39(1):btac817.

Xiang Dai, Sarvnaz Karimi, Ben Hachey, and Cecile Paris. 2020. An effective transition-based model for discontinuous ner. *arXiv preprint arXiv:2004.13454*.

Hao Fei, Donghong Ji, Bobo Li, Yijiang Liu, Yafeng Ren, and Fei Li. 2021. Rethinking boundaries: End-to-end recognition of discontinuous mentions with pointer networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 12785–12793.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

John Lafferty, Andrew McCallum, Fernando Pereira, et al. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Icml*, volume 1, page 3. Williamstown, MA.

Xiaonan Li, Hang Yan, Xipeng Qiu, and Xuanjing Huang. 2020. Flat: Chinese ner using flat-lattice transformer. *arXiv preprint arXiv:2004.11795*.


(c) 2024 中国中文信息学会计算语言学专业委员会

Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. 2022. Unified named entity recognition as word-word relation classification. In *proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 10965–10973.

Wei Liu, Xiyan Fu, Yue Zhang, and Wenming Xiao. 2021. Lexicon enhanced chinese sequence labeling using bert adapter. *arXiv preprint arXiv:2105.07148*.

Ling Luo, Zhihao Yang, Pei Yang, Yin Zhang, Lei Wang, Hongfei Lin, and Jian Wang. 2018. An attention-based bilstm-crf approach to document-level chemical named entity recognition. *Bioinformatics*, 34(8):1381–1388.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2019. Portuguese named entity recognition using bert-crf. *arXiv preprint arXiv:1909.10649*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Shanshan Wang, Yajing Yan, Rong Yan, Ting Li, Kaijie Ma, and Yani Yan. 2023. A chinese telemedicine-dialogue dataset annotated for named entities. *BMC Medical Informatics and Decision Making*, 23(1):264.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers*, pages 2335–2344.

Yue Zhang and Jie Yang. 2018. Chinese ner using lattice lstm. *arXiv preprint arXiv:1805.02023*.

Changmeng Zheng, Yi Cai, Jingyun Xu, HF Leung, and Guandong Xu. 2019. A boundary-aware neural model for nested named entity recognition. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics.

张帆 and 王敏. 2017. 基于深度学习的医疗命名实体识别. 计算技术与自动化, 36(1):123–127.

罗歆然, 李天瑞, and 贾真. 2024. 基于自注意力机制与词汇增强的中文医学命名实体识别. 计算机应用, 44:385–392.

CCL 2024

# 基于两阶段提示学习的少样本命名实体识别

邵佳兴　黄琪　肖聪　刘璟　罗文兵　王明文*
江西师范大学　计算机信息工程学院　江西　南昌　330022
Email: {shaojiaxing, huangqi, lwb, mwwang}@jxnu.edu.cn

## 摘要

少样本命名实体识别旨在用少量的标注数据来识别命名实体。近年来受提示学习在少样本场景中表现良好性能的启发，本文探索了基于提示的少样本命名实体识别的方法。已有的基于提示学习的方法是通过列举所有可能的跨度来进行实体识别，这导致了计算成本高以及对实体边界信息未充分利用的问题。本文提出一种基于提示学习的两阶段框架TSP-Few，在不使用源域数据的情况下，进行少样本命名实体识别。第一阶段对种子跨度进行增强、过滤和扩展，其中种子增强模块能够让种子跨度捕获到更丰富的语义信息，种子过滤器能够减少大量的无关跨度，种子扩展模块能够充分利用实体的边界信息，为实体类型分类提供高质量的候选实体跨度。第二阶段利用提示学习方法预测候选跨度的相应类别。此外，为了缓解跨度检测阶段的错误累积，在实体分类阶段引入了负采样策略。跨度检测和实体类型分类任务的独立训练更容易在少样本情况下取得优异的性能。在三个基准数据集上的实验表明，与先进的方法相比，本文提出的方法在性能上有了进一步的提升，并且实验结果也表明了该文模型各个模块的有效性。

**关键词：** 少样本学习 ； 命名实体识别 ； 提示学习

# Two-Stage Prompt Learning for Few-Shot Named Entity Recognition

**Jiaxing Shao　Qi Huang　Cong Xiao　Jing Liu　Wenbing Luo　Mingwen Wang***
School of Computer and Information Engineering, Jiangxi Normal University,
Nanchang, Jiangxi 330022,China
Email: {shaojiaxing, huangqi, lwb, mwwang}@jxnu.edu.cn

## Abstract

Few-Shot Named Entity Recognition (FS-NER) is designed to identify named entities using a limited amount of annotated data. Inspired by the effective performance of prompt-based learning in few-shot scenarios, This research examines methodologies for FS-NER that utilize prompting techniques. Conventional prompt-based methods involve enumerating all potential spans for entity recognition, which leads to high computational costs and insufficient use of boundary information about the entities. This

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版
**基金项目：** 国家自然科学基金(62266023); 江西省教育厅研究生创新基金项目(YC2023-s297);江西省教育厅科学技术研究项目(GJJ210325，GJJ2200354)
**通讯作者：** 王明文

paper proposes a two-stage framework based on prompt learning for FS-NER that operates without source domain data. In the first stage, seed spans are enhanced, filtered, and expanded: enhancement allows seed spans to capture richer semantic information, filtering eliminates a substantial number of irrelevant spans, and expansion utilizes the boundary information of entities effectively. The second stage employs prompt-based learning methods to predict the categories of the candidate spans. Moreover, to alleviate the accumulation of errors during the span detection phase, a negative sampling strategy is introduced in the entity classification stage. Independent training of span detection and entity type classification tasks proves to be more effective in few-shot settings. Experimental results on three benchmark datasets indicate that the proposed method shows further improvement over advanced methods, and the experiments also demonstrate the effectiveness of each module within the model.

**Keywords:** few-shot learning , named entity recognition , prompt learning

## 1 引言

命名实体识别（Named Entity Recognition，NER）作为一项基本的信息抽取任务，广泛应用于自然语言处理的下游任务。其目的是识别文本中的实体跨度，并将其分类为预先定义的实体类型类别，比如地名、组织名、时间、人名。尽管它具有挑战性，当前研究人员基于深度学习的方法已经取得了良好的性能 (Lample et al., 2016; Ma and Hovy, 2016; Chiu and Nichols, 2016; Peters et al., 2017)。然而，基于深度学习的方法依赖于大量的标注训练数据，这需要耗费大量时间和高度依赖于劳动密集度。除此之外，在一些特定的领域，需要特定领域知识，这是相当昂贵的。因此，近年来，在资源有限的情况下，进行命名实体识别，引起了学者们的广泛关注。

少样本命名实体识别的方法可以分为两类：基于跨度的方法和基于序列标注的方法。基于序列标注的方法通常是将命名实体识别任务看作是序列标注任务，给每句话中的单词都分配一个标签。 (Hou et al., 2020; Ballesteros et al., 2022)通过计算每个查询标记与支持标记之间的距离，在标记层面上直接对实体类型进行分类。然而，这些方法受到非实体标记（“O”类）和标记的标签依赖性的干扰。另一方面，基于度量学习的方法已经被应用到少样本命名实体识别。 (Snell et al., 2017)提出原型网络来学习一个度量空间，其中特定类别的一个查询例子聚集在一个原型周围。由于原型表示只是从几个例子中学习的，而other类往往是有非常多的噪声，所以无法在少样本命名实体识别达到令人满意的表现。

上述方法，都是从单词级别上进行命名实体识别，除此之外，命名实体识别任务还可以被视为跨度级别上的分类任务。大部分少样本命名实体识别均采用端到端度量学习的方法 (Huang et al., 2020),此类方法需要同时完成跨度检测和跨度分类两个任务。因此，在少样本场景下，仅依靠少数样本难以同时对两个任务充分学习，从而导致此类方法会产生实体边界预测错误，难以获得较好性能。 (Sohrab and Miwa, 2018)枚举出文本中所有的可能跨度，然后进行跨度分类。然而，这忽略了实体的边界信息，并且会产生大量的低质量的候选跨度，这严重影响模型性能。

最近，一些基于提示学习的工作被提出来用于少数命名实体识别。提示学习旨在通过将目标任务重新表述为完形填空问题 (Schick and Schütze, 2021)来提高预训练语言模型（Pretrained Language Model，PLM）的性能，这是PLM的预训练任务。例如，在执行情感分类任务时，使用带有[MASK]标记的模板将输入“我喜欢这部电影”转换为“我喜欢这部电影，它是[MASK]”。然后，PLM直接为[MASK]标记预测一个词“great”，并通过预测的词将输入归入正类。目标任务的转换减少了预训练和微调之间的差距，并有助于在少样本场景下工作。

为了解决基于跨度方法对实体边界信息利用不充分和枚举跨度带来的大量低质量的候选跨度的问题，受文献 (Yang et al., 2022)提出方法的启发，为了提高命名实体识别性能，本文提出了一种基于提示的两阶段框架，用于少样本命名实体识别。具体来说，在第一阶段，候选种子模块选择单词或双词作为初始种子，种子增强模块对初始种子进行语义增强，增强后的种子经过过滤，过滤掉低质量的种子。扩展模块是通过调整高质量种子的边界来定位实体，缩小了实


(c) 2024 中国中文信息学会计算语言学专业委员会

体类型预测的范围，提高候选跨度的质量。在第二阶段，使用提示学习方法，将候选跨度分类任务转化为掩蔽语言建模（MLM）问题。将实体类型预测建模为计算屏蔽位置的词汇概率分布问题。此外，跨度检测模块和实体类型分类模块的单一训练更能够在少样本情况下取得优异的性能。

本文的主要贡献有以下几点：

（1）为了提高命名实体识别性能，本文提出了一种基于提示学习的两阶段框架TSP-Few，用于少样本命名实体识别，它由跨度检测模块和实体类型分类模块组成。在跨度检测阶段，为了增强初始种子的语义信息，设计了一个种子增强模块。

（2）为了缓解将少样本命名实体识别分解为跨度检测和实体类型分类所造成的错误累积，本文将负向采样策略纳入到模型中。

（3）在三个基准数据集上的不同设置下的实验表明，与最先进的方法相比，本文提出的方法的性能取得了明显的改善。

## 2 相关工作

少样本命名实体识别任务旨在少样本场景下对命名实体跨度进行识别并分类为特定的实体类别。以前关于少样本命名实体识别的研究大多是基于度量学习的。具体地，(Yi and Katiyar, 2020)等人提出Structshot算法，该算法在源域上训练命名实体识别模型作为特征提取器，推理时利用最近邻算法对实体类型分类并使用维特比解码器学习标签的依赖性。(Fritzler et al., 2019)首先将原型网络调整应用于少样本命名实体识别，然后 (Tong et al., 2021)等人通过训练分类器学习多个原型表示，用于表示“O”类嘈杂的语义，弥补了现有的原型方法不能区分“O”类词中丰富的语义信息的不足。然而，在少样本场景下令牌级基于度量学习的方法难以充分学习两个不同的任务。

为了解决基于度量学习方法存在的问题，有研究者提出基于跨度的方法用于少样本命名实体识别，(Ma et al., 2022)将命名实体识别任务分解为跨度检测和实体分类，将跨度预测当作序列标注问题并使用MAML算法训练跨度检测器，在实体分类，利用MAML算法增强原型网络，找到好的嵌入空间来更好的分辨不同实体类的跨度。最近，(Wang et al., 2015)等人提出ESD模型，该模型利用多个不同注意力机制对支持集和查询集之间的跨度信息进行交互，并且使用跨度级原型网络，利用查询集跨度对原型进行增强，同时，将“O”类型跨度分为3个子类，以缓解其复杂的语义问题。(Wang et al., 2022)等人在跨度提取阶段，将顺序标签转化为全局边界矩阵，使模型专注于明确的边界信息，并提出基于边缘的损失来缓解两阶段方法中假阳性的问题，并且利用原型学习在语义空间中调整跨度的表示。

随着GPT (Brown et al., 2020)的引入，提示学习被广泛讨论，并在少量学习中显示出巨大的性能 (Gao et al., 2021) 。最近的一些研究试图引入提示学习，并使其适应少样本命名实体识别。(Cui et al., 2021)等人是第一个提出使用生成式预训练语言模型来解决少样本序列标注问题，通过提示模板列举输入文本的所有候选跨度，模型根据相应的模板分数对每个候选跨度进行分类。(Ma et al., 2021)等人提出了一种无模板的方法，对每个输入只使用一个预测，这个模型减少了预训练和微调之间的差异，并且没有修改预训练模型的输出层。(Hou et al., 2022)等人为了提高基于提示学习预测速度，提出了一种反向的提示范式，在给定槽位标签类型的情况下，反向预测槽位值，而不是列举槽位值来预测标签。

然而，这些研究大多同时学习跨度检测和类型分类任务，这在少样本场景下很难取得优异的表现。此外，这些方法遵循情节训练的方法，或者假设有一个丰富的源域，这不是一个更实用的少样本场景。为了处理上述问题，本文提出了一种基于提示的两阶段少样本命名实体识别框架，它由一个跨度检测模块和一个实体类型分类模块组成。

## 3 模型

本文提出一种基于提示的两阶段框架TSP-Few，用于少样本命名实体识别（Few-shot NER）。如图1所示，该框架由两个预先训练好的语言模型组成，其中一个用于解决实体跨度检测，输入文本首先经过候选种子模块产生候选种子，然后候选种子进入种子增强模块，对种子进行增强，之后通过计算出的阈值过滤出高质量的候选种子，最后经过种子扩展得到候选跨度。跨度检测模型是与类别无关的，可以在不同的实体类别之间共享信息。另一个用于候选跨度分类，本文将经过扩展得到的候选跨度填入预先定义的模板并送入实体分类模型，该模型缩


（c）2024 中国中文信息学会计算语言学专业委员会

小了实体类型预测的范围，旨在通过提示学习为每个跨度分配一个特定的实体类别。下面我们分别讨论每个模块的细节。

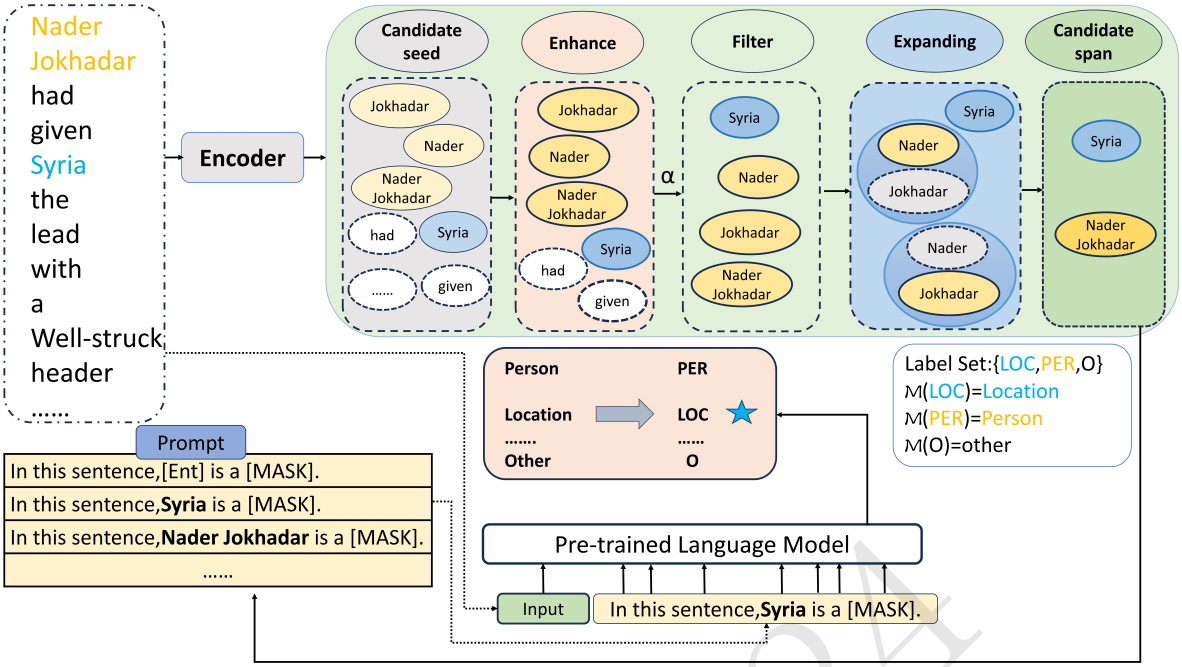



图 1: 模型框架

### 3.1 问题陈述

本文将命名实体识别任务分解为两个子任务：跨度检测和实体类型分类。给定一个长度为n的输入文本$X = \{x_1, ...x_n\}$，跨度检测的目的是提取每个候选跨度$s = [x_i, ...., x_j](1 \leq i \leq j \leq n)$，并从输入文本中获得一个候选跨度集$S = \{s_1, ...s_m\}$，实体类型分类模块的目标是为每个候选跨度$s \in S$分配一个实体类型$y \in Y$。

### 3.2 第一阶段

#### 3.2.1 产生候选种子

候选种子是从输出文本中抽取的子序列。通过对过滤之后的候选种子进行扩展，能够得到候选跨度。输入文本$X = \{x_1, ..., x_n\}$，列举出长度L不超过2的所有跨度。把得到的候选种子跨度的集合表示为$C = \{c_1, ..., c_{2n-1}\}$，其中$c_i = (st_i, ed_i)$表示第$i$个候选种子跨度，$st_i$、$ed_i$分别表示该跨度的左边界和右边界。这里使用BERT作为编码器，将输入文本送入BERT，获得隐藏表示$h \in R^{n \times d}$，这里d是BERT的隐藏层状态的维度。

为了训练候选种子过滤器，筛选出高质量的候选种子，这更有可能扩展为正确的命名实体词，需要计算出候选种子的分数。具体来说，本文将$C$中的每个候选种子跨度和真实实体配对，计算出候选种子分数。计算候选种子分数一般有两种方式：（1）Intersection over Foreground (IoF)，（2）Intersection over Union (IoU)。在命名实体识别中，IoU被定义为$IoU(A, B) = \frac{A \cap B}{A \cup B}$，其中A、B是两个跨度 (Zheng et al., 2019; Shen et al., 2021)。IoU衡量跨度之间的重叠程度，这对嵌套命名实体识别更有作用。然而，这里很少涉及嵌套命名实体，所以本文使用IoF计算候选种子分数。在命名实体识别中，IoF被定义为$IoF(A, B) = \frac{A \cap B}{A}$。(Yang et al., 2022)考虑到一个由五个词组成的命名实体，“International Conference on Computational Linguistics”，候选种子为“International Conference”与该实体之间的IoU为0.4，并不显著。而该候选种子与实体之间的IoF为1，其中，A是候选种子，B是命名实体。这表明该候选种子是实体的一部分，很有可能扩展到整个实体。这里，使用IoF计算候选种子得分更加合适。计算每个候选种子与其真实实体之间的IoF，作为候选种子分数$y^{cw}$。


(c) 2024 中国中文信息学会计算语言学专业委员会

### 3.2.2 种子增强模块

本文认为，种子跨度的含义能够从其他种子跨度中推断出来。因此，本文设计了一个种子注意力机制。给出一个句子$X \in R^{K\times d}$($K$是跨度的数量)的所有跨度表示，对于候选种子跨度$c_i = (st_i, ed_i)$，$h_i^p$代表跨度$c_i$的平均池化表示。对于种子$c_i$,得到它的增强跨度表示$\bar{h_i^p} = \sum_{j=1}^{K}\alpha_i^j c_j$,这里$\alpha_i = softmax(c_i X^T)$, 然后，使用带有残差连接和层归一化的前馈神经网络来获得最终的种子增强表示$\hat{h_i^p}$,

$$h_i^p = MeanPooling(h_{st_i}, ..., h_{ed_i}) \tag{1}$$

$$\hat{h_i^p} = LayerNorm(h_i^p + FFN(\bar{h_i^p})) \tag{2}$$

$$FFN(\bar{h_i^p}) = GELU(\bar{h_i^p}W^1)W^2 \tag{3}$$

其中，$W^1$、$W^2$是权重矩阵，这里省略了偏置项。

候选种子过滤器：能够计算出所有候选种子的得分，并设置阈值$\alpha$，选择得分大于$\alpha$的候选种子进行扩展。对于候选种子跨度$c_i = (st_i, ed_i)$，$\hat{h_i^p}$代表经过种子增强的跨度表示，连接$\hat{h_i^p}$和[CLS]标记的$h^{CLS}$表示作为候选种子跨度的最终表示$h_i^{cw}$。候选种子分数$p_i^{cw}$的计算方法如下:

$$h_i^{cw} = Concat(\hat{h_i^p}, h^{[CLS]}) \tag{4}$$

$$p_i^{cw} = Sigmoid(MLP(h_i^{cw})) \tag{5}$$

这里MLP表示多层感知机，最后一层使用GELU激活函数。

### 3.2.3 候选种子扩展

尽管通过种子过滤器得到了质量较高的候选种子，但这些种子还不能够精准的定位命名实体，还需要对候选种子边界进行调整，这是一项回归任务。为了学习边界偏移量$\hat{o}$不仅需要跨度本身的信息，也需要外部边界词的信息。因此，扩展还需要考虑候选种子周围的一个窗口。

扩展是允许候选种子的左右边界分别偏移$\lambda$，这样能够得到最长的实体的长度是$2+2\lambda$，对于候选种子$c_i$，最大扩展为:

$$b_i^{max} = (max(1, st_i - \lambda), min(n, ed_i + \lambda)) \tag{6}$$

为了能够区分最大扩展的边界，窗口应该大于$b_i^{max}$，定义为$w_i$。

$$w_i^l = max(1, st_i - 2\lambda) \tag{7}$$

$$w_i^r = min(n, ed_i + 2\lambda) \tag{8}$$

$$w_i = (w_i^l, w_i^r) \tag{9}$$

其中，$w_i^l$，$w_i^r$分别表示窗口的左右边界。

将候选种子$c_i$的平均集合跨度$h_i^p$和窗口的平均集合跨度$h_i^w$连接起来，然后计算边界偏移量$o_i$，计算如下:

$$h_i^w = MeanPooling(h_{w_i^l}, ..., h_{w_i^r}) \tag{10}$$

$$o_i = \lambda \cdot (2 \cdot Sigmoid(MLP(h_i^p; h_i^w)) - 1) \tag{11}$$

其中，（;）表示连接操作。$o_i$的第一个元素表示为$o_i^{st}$，第二个元素表示为$o_i^{ed}$。$o_i^{st}$表示为左边界偏移量，$o_i^{ed}$表示右边界偏移量。这样，就可以得到新的左右边界的候选跨度。

$$l_i^{'} = max(1, st_i + \left\lfloor o_i^{st} + \frac{1}{2} \right\rfloor) \tag{12}$$

$$r_i^{'} = max(n, ed_i + \left\lfloor o_i^{ed} + \frac{1}{2} \right\rfloor) \tag{13}$$

其中$l_i^{'}$，$r_i^{'}$分别表示新的左右边界。重复的结果和无效的结果$l_i^{'} > r_i^{'}$会被丢弃。这样，产生了一组候选跨度，用于跨度分类。


(c) 2024 中国中文信息学会计算语言学专业委员会

#### 3.2.4 训练目标

产生候选种子损失$L_{cw}$和候选种子扩展损失$L_{exp}$定义为：

$$L_{cw} = \sum_{i} SmoothL1(y_i^{cw}, p_i^{cw}) \tag{14}$$

$$L_{exp} = \sum_{i} \sum_{j \in \{st,ed\}} SmoothL1(\hat{o}_i^{y}, o_i^{j}) \tag{15}$$

其中，$y_i^{cw}$是真实的候选种子分数，$\hat{o}_i^{y}$是原始的边界偏移量。

通过最小化综合损失$L_z$进行训练，定义如下：

$$L_z = L_{cw} + L_{exp} \tag{16}$$

### 3.3 第二阶段

在实体类型分类模块中，预测由上述跨度检测模块从候选种子扩展得到的候选跨度集$S = \{s_1, ..., s_m\}$中每个候选跨度的实体类别。为避免引入新的参数并减少预训练和微调之间的差距，本文遵循基于提示的调整，将分类任务转化为掩蔽语言建模（MLM）问题。基于提示的方法包括构建言语器和模板。在接下来的章节中，将详细介绍实体类型预测模块的言语器和模板的构建。

#### 3.3.1 言语器设计

言语器的作用是将标签词汇映射到一个特殊的类别，这有助于预训练语言模型理解每个类别。为了丰富实体类别的内涵，实体类型分类模块的言语器设计是将每个实体类别$y \in Y$映射到一个词$\nu_y$，即$M : y \rightarrow \nu$。例如，标签集合$Y = \{PER, LOC, ORG, O\}$被转换为相应的词$\nu = \{person, location, organization, other\}$。

#### 3.3.2 模板设计

为了统一预训练任务和微调任务的形式，提示模板将原始输入$x$修改为提示输入$T(x)$,其中包括[MASK]标记。在本文的框架中，用自然语言构建一个模板$T(x)$，如下所示：

$$T(x) = x.In\ this\ sentence, [Ent]\ is\ a\ [MASK]. \tag{17}$$

其中$x$代表原始句子输入，$[Ent]$表示句子$x$中的候选实体跨度。因此，计算标签$y \in Y$概率的实体类型预测任务被建模为计算被屏蔽位置的词汇的概率分布，其形式化为：

$$p(y|x) = p([MASK] = M(y)|T(x)) \tag{18}$$

#### 3.3.3 训练和推理

为了缓解两阶段框架带来的误差累积问题，本文采用了负向采样策略，在训练过程中对训练数据进行采样。具体来说，随机抽取非实体的跨度作为实体类型分类的负面样本，并利用“$O$”来表示负面样本$O \in Y$的类型。采用负面抽样策略使实体分类模块有能力预测非实体类别。当跨度检测模块出现错误时，实体分类模块能够将错误跨度分配“other”，从而缓解了错误的积累。此外，在实体分类模块中，本文选择BERT作为预训练模型，并利用交叉训练损失作为目标函数来微调预训练模型。该函数的形式如下：

$$\mathcal{L} = -\sum \log p(y|x) \tag{19}$$

最后，将跨度检测模块得到的候选跨度填入$T(x)$中的$[Ent]$，并将其送入微调后的预训练模型，在推理过程中预测跨度的实体类型。

第二十三届中国计算语言学大会论文集，第394页-第405页，太原，中国，2024年7月25日至28日。
卷1：主会论文
(c) 2024 中国中文信息学会计算语言学专业委员会

# 4 实验设置

## 4.1 少样本设置

与之前假设有源域数据可用的少样本命名实体识别不同，本文考虑了没有源域训练的设置，这更具有挑战性。这个实际场景只有一个测试集，每个类别只有N个标签（N-way）的K个例子（K-shot）。为了更好的评估基线和本文提出的模型的性能，本文用$K \in \{5, 10, 20\}$进行了实验。

## 4.2 数据集的构建

为了给所提出的模型提供有效性的经验证据和为了进行公平的比较，本文在三个数据集上进行了实验。在$K \in \{5, 10, 20\}$的情况下，本文完全遵循 (Yang and Katiyar, 2020)中的贪心抽样策略。具体来说，（1）训练集中的每个类别至少有K个对应的命名实体。（2）从训练集中删除任何句子都会导致至少有一个类别的命名实体少于K。为了保证实验的有效性和可靠性，所使用的三个数据集中的抽样数据是由 (Yang et al., 2022)所提供公开的。本文使用的三个不同领域的命名实体识别数据集分别是：CoNLL03数据集 (Sang and Meulder, 2003)在新闻领域、MIT-Restaurant数据集 (Liu et al., 2013)在评论领域和WiKiGold数据集 (Balasuriya et al., 2009)一般领域。表1显示了这些数据集的原始信息。

| Dataset | Domain | Language | Class | Train | Test |
|---|---|---|---|---|---|
| CoNLL03 | News | English | 4 | 14,987 | 3,684 |
| MIT-Restaurant | Review | English | 8 | 7,660 | 1,521 |
| WikiGold | General | English | 4 | 1,017 | 339 |

表 1: 原始数据集的统计信息

## 4.3 实施细节

本文使用PyTorch (Paszke et al., 2019)和Openprompt (Paszke et al., 2022)实现模型。在跨度检测模块，英文数据集使用"bert-base-uncased"预训练模型作为编码器，epochs设置为35，CoNLL03、MIT-Restaurant和WiKiGold数据集的阈值$\alpha$分别设置为0.5，0.6，0.7。$\lambda$设置为5。采用AdamW优化器，学习率为3e-0.5。对于实体类型分类模块，本文使用"bert-base-uncased" 预训练模型作为编码器，并设置epochs的数量为30，该模块使用AdamW作为优化器，学习率为5e-5。对于所有的基线，本文使用原始论文或者官方实现提供的推荐参数。

## 4.4 基线

本文将提出的模型与六个竞争性基线进行比较。

Prototype (Snell et al., 2017)一种基于原型网络的方法，将实体类别表示为单个标记的同一表示空间中的向量，并利用最近的邻居标准来分配实体类别。

LC-BERT (Devlin et al., 2019) 一个基于BERT 的序列标签模型，它直接在BERT模型中加入一个线性分类器，用于序列标签。

NNShot和StructShot (Yang and Katiyar, 2020)提出了一个基于度量的少样本学习方法NNShot 用于命名实体识别，它采用最近邻分类器进行预测。在此基础上，他们进一步提出了在解码过程中引入维特比算法的少样本学习方法StructShot。

TemplateNER (Cui et al., 2021)一种基于模板的提示学习方法，用于少样本命名实体识别，它在推理过程中列举所有的跨度来生成预定义的模板作为BART 的输入。

SEE-Few (Yang et al., 2022)将命名实体识别分为两部分，跨度检测和跨度分类。在跨度检测模块生成种子，直接经过筛选，再对种子扩展为候选跨度。对于跨度分类模块，它将跨度分类重新定义为文本蕴含任务，同时利用上下文和实体类型信息进行类型预测。

# 5 实验结果

## 5.1 结果分析

表2显示了本文所提出的方法和基线在CoNLL03、MIT-Restaurant和WiKiGold数据集


(c) 2024 中国中文信息学会计算语言学专业委员会

上不同少样本设置中的表现，有以下重要观察：1）本文的模型在上述三个数据上表现出优越性。具体来说，在K=5样本设置下，本文提出的模型的F1分数在CoNLL03、MIT-Restaurant和WiKiGold 比基线模型分别增加了2.27%、12.75%、6.77%；在K=10样本设置下分别增加了：6.91%、13.24%、3.94%，这验证了本文提出的模型在少样本设置下表现更加优异，对少样本数据下的情况更加适应。2）与基线相比，本文提出的模型能够在5样本设置下取得比基线在20样本设置下表现相当甚至更优的性能表现。3）本文的方法优于以前的提示方法，与TemplateNER模型在不同样本设置对比下更加明显。4）在本文的少样本设置中，优于缺乏丰富的源域数据进行预训练，基于度量的方法NNShot和StructShot并没有显示出它们的优势，这些结果表明这些方法不适合更实际的少样本场景。

| Datasets | Methods | K=5 | | | K=10 | | | K=20 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | P | R | F1 | P | R | F1 | P | R | F1 |
| **CoNLL03** | LC-BERT | 42.83 | 30.72 | 35.06$_{(6.09)}$ | 50.36 | 52.13 | 51.20$_{(6.39)}$ | 56.33 | 63.85 | 59.84$_{(1.43)}$ |
| | Prototype | 38.26 | 43.14 | 40.37$_{(8.06)}$ | 45.08 | 64.02 | 52.82$_{(3.22)}$ | 43.94 | 69.72 | 53.89$_{(1.95)}$ |
| | NNShot | 32.11 | 38.42 | 34.92$_{(3.30)}$ | 34.10 | 40.98 | 37.18$_{(5.82)}$ | 38.43 | 47.85 | 42.61$_{(2.23)}$ |
| | StructShot | 30.04 | 21.33 | 23.43$_{(4.52)}$ | 38.62 | 19.72 | 26.09$_{(7.23)}$ | 44.96 | 28.59 | 34.87$_{(1.30)}$ |
| | TemplateNER | 26.90 | 23.46 | 23.13$_{(8.40)}$ | 44.51 | 43.99 | 44.01$_{(4.82)}$ | 52.16 | 56.46 | 54.01$_{(5.09)}$ |
| | SEE-Few | 60.45 | 51.27 | 55.21$_{(3.93)}$ | 66.19 | 58.68 | 61.99$_{(1.73)}$ | 69.49 | 67.07 | 68.21$_{(2.60)}$ |
| | Ours | **63.63** | **52.42** | **57.48$_{(5.10)}$** | **70.82** | **67.08** | **68.90$_{(1.94)}$** | **76.89** | **76.59** | **76.74$_{(3.17)}$** |
| **MIT-Restaurant** | LC-BERT | 41.21 | 38.65 | 39.88$_{(3.79)}$ | 43.60 | 48.93 | 46.08$_{(3.75)}$ | 56.24 | 60.04 | 58.07$_{(1.50)}$ |
| | Prototype | 27.77 | 46.79 | 34.84$_{(1.63)}$ | 30.37 | 50.64 | 37.97$_{(2.29)}$ | 37.91 | 59.31 | 46.25$_{(1.62)}$ |
| | NNShot | 28.15 | 34.81 | 31.11$_{(2.30)}$ | 30.28 | 37.56 | 33.56$_{(1.48)}$ | 36.72 | 45.55 | 40.66$_{(1.26)}$ |
| | StructShot | 45.13 | 25.00 | 31.93$_{(4.32)}$ | 43.94 | 28.19 | 34.30$_{(2.56)}$ | 52.08 | 36.18 | 42.69$_{(1.12)}$ |
| | TemplateNER | 23.11 | 20.78 | 21.53$_{(4.66)}$ | 39.45 | 28.77 | 32.71$_{(8.14)}$ | 46.93 | 37.00 | 41.26$_{(6.80)}$ |
| | SEE-Few | 53.08 | 39.47 | 45.25$_{(3.18)}$ | 57.19 | 46.41 | 51.20$_{(1.48)}$ | 64.79 | 57.22 | 60.75$_{(2.07)}$ |
| | Ours | **71.34** | **49.85** | **58.69$_{(3.44)}$** | **74.61** | **56.71** | **64.44$_{(2.15)}$** | **76.17** | **68.16** | **71.94$_{(1.88)}$** |
| **WikiGold** | LC-BERT | 36.02 | 8.02 | 12.57$_{(7.81)}$ | 43.13 | 8.95 | 37.72$_{(7.20)}$ | 50.68 | 50.73 | 50.68$_{(5.94)}$ |
| | Prototype | 20.55 | 21.46 | 19.28$_{(8.12)}$ | 23.31 | 45.21 | 30.59$_{(3.95)}$ | 27.31 | 56.22 | 36.56$_{(8.65)}$ |
| | NNShot | 27.81 | 34.16 | 30.63$_{(1.91)}$ | 26.36 | 37.92 | 30.93$_{(4.89)}$ | 28.33 | 39.07 | 32.81$_{(5.41)}$ |
| | StructShot | 49.00 | 13.37 | 20.88$_{(4.61)}$ | 43.21 | 14.19 | 21.28$_{(2.96)}$ | 43.51 | 15.94 | 23.16$_{(2.18)}$ |
| | TemplateNER | 18.45 | 19.45 | 17.26$_{(12.73)}$ | 38.33 | 45.37 | 41.04$_{(13.19)}$ | 57.39 | 56.00 | 56.00$_{(3.22)}$ |
| | SEE-Few | 61.23 | 41.01 | 48.87$_{(8.01)}$ | 63.36 | 48.74 | 54.98$_{(3.24)}$ | 69.06 | 58.25 | 63.19$_{(1.28)}$ |
| | Ours | **74.94** | **44.24** | **55.64$_{(5.84)}$** | **67.19** | **58.46** | **58.92$_{(5.80)}$** | **76.58** | **64.52** | **70.03$_{(2.65)}$** |

表 2: 三个数据集上不同场景下的实验结果

## 5.2 消融实验

为进一步验证所提方法中不同组件的有效性，本文用不同的少样本设置进行了消融实验。本文构建了多个消融实验变体：

1）w/o sea：移除跨度检测模块中的种子增强模块，直接将种子进行过滤；

2）w/o neg：移除跨度分类模块中的负向采样策略；

3）w/o span：移除跨度检测模块，在CoNLL03、MIT-Restaurant和WiKiGold数据集上列举长度L=5的所有可能跨度，并将其填入实体分类模块；

4）w/o prompt：移除提示学习模块，直接使用基线SEE-Few模型的跨度分类方法，将分类任务定义为文本蕴含任务。

| Model | CoNLL03 | | | MIT-Restaurant | | | WikiGold | | |
|---|---|---|---|---|---|---|---|---|---|
| | 5 | 10 | 20 | 5 | 10 | 20 | 5 | 10 | 20 |
| Full model | **57.48** | **68.90** | **76.74** | **58.98** | **64.57** | **71.94** | **55.64** | **58.92** | **70.03** |
| w/o sea | 54.72 | 65.06 | 75.50 | 58.21 | 60.84 | 70.67 | 52.06 | 56.30 | 68.39 |
| w/o neg | 54.98 | 67.10 | 76.24 | 57.91 | 64.26 | 71.18 | 54.78 | 57.69 | 69.73 |
| w/o span | 38.14 | 49.85 | 55.25 | 35.71 | 42.44 | 38.37 | 27.12 | 17.80 | 20.12 |
| w/o prompt | 56.05 | 55.88 | 68.42 | 49.92 | 51.86 | 59.49 | 50.30 | 45.85 | 66.77 |

表 3: 在不同的样本设置下，模型消融实验结果在三个数据集上的F1得分

本文对消融实验采用F1值作为实验结果，如表3所示。由实验结果可知：


(c) 2024 中国中文信息学会计算语言学专业委员会

| Model | CoNLL03 | MIT-Restaurant | WiKiGold |
|---|---|---|---|
| Full model | 5648 | 3151 | 730 |
| w/o span | 199068 | 56139 | 32073 |

表 4: 测试集上候选跨度的数量统计

（1）对于移除种子增强模块，直接将种子进行过滤的设置。这意味着种子跨度的含义不能从其他种子跨度中推断出来，实验结果表明，在CoNLL03、MIT-Restaurant、WiKiGold数据集上5样本设置下，模型的性能F1分数分别下降了2.76%、0.77%、3.58%，这表明，种子增强模块能够让特定的种子跨度从其他种子跨度推断出来，证明种子增强模块是有效的。

（2）对于移除负向采样策略，模型不能够学习到“other”类别的信息，从实验结果来看，在所有的数据集和少样本设置中，取消负向采样策略后，F1分数都在下降。这表明，利用负向采样有助于缓解错误累积，能够识别出跨度检测阶段识别出的虚假跨度。

（3）对于移除跨度检测模块，直接枚举跨度，进行分类。可以看到，与移除负采样相比，跨度检测模块对少样本命名实体识别有更大的影响，直接枚举跨度导致模型性能急剧下降。从表4的实验结果分析，这是因为直接枚举跨度会产生数量巨大的无关跨度，这些杂乱的语义信息对模型来说是灾难性的，这是导致模型性能下降的直接原因。这表明，跨度检测模块能够极大的减少大量的无关跨度，这对模型性能很重要，跨度检测模块有能力缩小命名实体识别的范围并提高性能。

（4）对于移除提示学习模块，将跨度分类任务定义为文本蕴含任务。从实验结果来看，F1分数都有明显下降，这说明利用提示学习进行实体类型分类，模型能够更好的充分利用上下文的信息，能够更准确的识别实体类型。除此之外，模型在K=10的样本设置下，性能下降的更快，根据实验结果分析，在K=10的场景中，在第一阶段中的种子增强模块，可能给种子跨度引入了噪声，并没有对所有的种子跨度都有增强效果，反而影响最终的模型性能。

### 5.3 案例研究

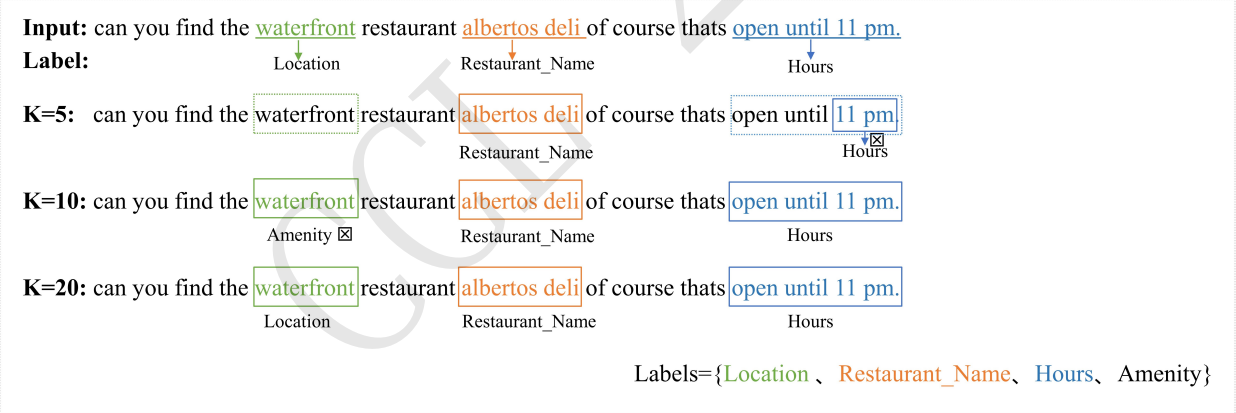



图 2: 案例研究。在数据集MIT-Restaurant下不同样本场景中进行的案例分析。实线框表示模型识别到的实体跨度，虚线框表示模型未识别到的实体跨度。

图2展示了在不同样本设置下，模型预测的例子。我们可以观察到在K=5样本设置下，模型只正确检测出实体跨度“albertos deli”，而模型没有检测出实体跨度“waterfront”，并且模型检测出的错误跨度“11 pm”只是真实实体跨度“open until 11 pm”的一部分，没有扩展成完整的实体跨度，原因是：在K=5样本设置下，训练样本信息太少，对种子扩展模块还不足以充分的训练。在K=10样本设置下，模型正确检测出实体跨度“waterfront”和“open until 11 pm”，然而，模型虽然能检测出在K=5样本中未识别出的跨度，但是，在对实体跨度“waterfront”进行分类时，模型分配了错误的实体类型“Amenity”，而正确的实体类型是“Location”。我们分析认为是模型对实体类型的信息捕获不足及对上下文信息的利用还不够充分。此外，模型能够准确检测


(c) 2024 中国中文信息学会计算语言学专业委员会

出实体跨度“open until 11 pm”，这表明模型具备将种子跨度扩展到完整实体跨度的能力。最后，在K=20样本设置下，模型不仅能够准确检测出实体跨度还能够为实体跨度分配准确的实体类型。这说明模型在训练样本数增加下，能够学习到更多的知识，取得较好的性能。

## 6 总结与未来工作

在本文中，提出一个两阶段框架TSP-Few，用于不使用源域数据的少样本命名实体识别。本文将少样本命名实体识别任务分解为跨度检测和实体分类两个阶段，这可以使模型在少样本情况下取得更出色的性能。种子增强模块和扩展模块对实体分类模块提供尽可能准确的候选跨度。实体分类模块将分类任务转化为屏蔽语言建模任务，将实体类型分配给候选跨度，通过减少预训练和微调之间的差距来提高少样本性能。在三个基准数据集上的大量实验表明，本文提出的模型在少样本命名实体识别基准上取得了有竞争力的结果。

在未来的工作中，我们将探索把提示学习与对比学习结合起来，以提高PLM的性能，并进一步提高少样本命名实体识别的性能。

## 参考文献

Dominic Balasuriya, Nicky Ringland, Joel Nothman, Tara Murphy, and James R. Curran. 2009. Named Entity Recognition in Wikipedia. In *Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources (People's Web)*, pages 10–18.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20)*, pages 1877-1901.

Jason P.C. Chiu and Eric Nichols. 2016. Named Entity Recognition with Bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, pages 4:357–370.

Leyang Cui, Yu Wu, Jian Liu, Sen Yang, and Yue Zhang. 2021. Template-Based Named Entity Recognition Using BART. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1835–1845.

Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *North American Chapter of the Association for Computational Linguistics(2019).*

Ning Ding, Shengding Hu, Weilin Zhao, Yulin Chen, Zhiyuan Liu, Haitao Zheng, and Maosong Sun. 2022. OpenPrompt: An Open-source Framework for Prompt-learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 105–113.

Alexander Fritzler, Varvara Logacheva, and Maksim Kretov. 2019. Few-shot classification in named entity recognition task. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing (SAC '19)*, pages 993-1000.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making Pre-trained Language Models Better Few-shot Learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830.

Jiaxin Huang, Chunyuan Li, Krishan Subudhi, Damien Jose, Shobana Balakrishnan, Weizhu Chen, Baolin Peng, Jianfeng Gao and Jiawei Han. 2020. Few-Shot Named Entity Recognition: A Comprehensive Study. *arXiv preprint arXiv:2012.14978 (2020).*

Yutai Hou, Cheng Chen, Xianzhen Luo and Bohan Li. 2022. Inverse is better! fast and accurate prompt for few-shot slot tagging. *arXiv preprint arXiv:2204.00885 (2022).*


(c) 2024 中国中文信息学会计算语言学专业委员会

Yutai Hou, Wanxiang Che, Yongkui Lai, Zhihan Zhou, Yijia Liu, Han Liu, and Ting Liu. 2020. Few-shot Slot Tagging with Collapsed Dependency Transfer and Label-enhanced Task-adaptive Projection Network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1381–1393.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural Architectures for Named Entity Recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270.

Jingjing Liu, Panupong Pasupat, D. Scott Cyphers and James R. Glass. 2013. Asgard: A portable architecture for multilingual dialogue systems. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing(2013)*, pages 8386-8390.

Jie Ma, Miguel Ballesteros, Srikanth Doss, Rishita Anubhai, Sunil Mallya, Yaser Al-Onaizan, and Dan Roth. 2022. Label Semantics for Few Shot Named Entity Recognition. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1956–1971.

Ruotian Ma, Xin Zhou, Tao Gui, Yi Chern Tan, Qi Zhang and Xuanjing Huang. 2021. Template-free Prompt Tuning for Few-shot NER. *arXiv preprint arXiv:2109.13532 (2021).*

Tingting Ma, Huiqiang Jiang, Qianhui Wu, Tiejun Zhao, and Chin-Yew Lin. 2022. Decomposed Meta-Learning for Few-Shot Named Entity Recognition. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1584–1596.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in neural information processing systems 32 (2019).*

Matthew E. Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Powe. 2017. Semi-supervised sequence tagging with bidirectional language models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1756–1765.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Mohammad Golam Sohrab and Makoto Miwa. 2021. Deep Exhaustive Model for Nested Named Entity Recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2843–2849.

Jake Snell,Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*, pages 4080–4090.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*, pages 4080–4090.

Timo Schick and Hinrich Schütze. 2021. Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 255–269.

Yongliang Shen, Xinyin Ma, Zeqi Tan, Shuai Zhang, Wen Wang, and Weiming Lu. 2021. Locate and Label: A Two-stage Identifier for Nested Named Entity Recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*,pages 2782–2794.

Meihan Tong, Shuai Wang, Bin Xu, Yixin Cao, Minghui Liu, Lei Hou and Juan-Zi Li. 2021. Learning from Miscellaneous Other-Class Words for Few-shot Named Entity Recognition. *Annual Meeting of the Association for Computational Linguistics(2021).*


(c) 2024 中国中文信息学会计算语言学专业委员会

Jianing Wang, Chengyu Wang, Chuanqi Tan, Minghui Qiu, Songfang Huang, Jun Huang, and Ming Gao. 2022. SpanProto: A Two-stage Span-based Prototypical Network for Few-shot Named Entity Recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3466–3476.

Peiyi Wang, Runxin Xu, Tianyu Liu, Qingyu Zhou, Yunbo Cao, Baobao Chang, and Zhifang Sui 2015. An Enhanced Span-based Decomposition Method for Few-Shot Sequence Labeling. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5012–5024.

Yang, Yi, and Arzoo Katiyar. 2020. Simple and Effective Few-Shot Named Entity Recognition with Structured Nearest Neighbor Learning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2540–2550.

Yi Yang and Arzoo Katiyar. 2020. Simple and Effective Few-Shot Named Entity Recognition with Structured Nearest Neighbor Learning. *arXiv preprint arXiv:2010.02405 (2020)*.

Zeng Yang, Linhai Zhang, and Deyu Zhou. 2022. SEE-Few: Seed, Expand and Entail for Few-shot Named Entity Recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2540–2550.

Changmeng Zheng, Yi Cai, Jingyun Xu, Ho-fung Leung, and Guandong Xu. 2019. A Boundary-aware Neural Model for Nested Named Entity Recognition. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 357–366.

CCL 2024

# 面向工艺文本的实体与关系最近邻联合抽取模型

**杨丹清忻　王裴岩*　徐立军**
沈阳航空航天大学计算机学院，辽宁沈阳110134
wangpy@sau.edu.cn
yangdanqingxin@stu.sau.edu.cn
Xlj@ge-soft.com

## 摘要

该文研究工艺文本中实体关系联合抽取问题，提出了最近邻联合抽取模型（NNJE）。NNJE利用工艺文本中实体边界字间搭配规律建模外显记忆，通过最近邻方法在某种指定关系下为待预测组合检索出具有相似字间搭配的实例，为实体边界识别以及实体对组合提供更有力的限制条件，提升模型预测准确率，改善模型性能。实验设置了工艺文本关系数据集。实验结果表明，该文方法较基线模型准确率P值提高了3.53%，F1值提升了1.03%，优于PURE、CasRel、PRGC与TPlinker等方法，表明提出的方法能够有效地提升三元组抽取效果。

**关键词：** 工艺文本；实体关系联合抽取；最近邻

# Nearest Neighbor Joint Extraction Model for Entity and Relationship in Process Text

**Danqingxin Yang　Peiyan Wang　Lijun Xu**
School of Computer Science, Shenyang Aerospace University,
Shenyang, Liaoning 110134, China
wangpy@sau.edu.cn
yangdanqingxin@stu.sau.edu.cn
Xlj@ge-soft.com

## Abstract

To address the issue that joint extraction of entity and relationships in process texts, this paper proposes the Nearest Neighbor Integrated Extraction model (NNJE). NNJE uses the rules of word combinations in the process text to model explicit memory and retrieves combinations with similar collocation for the predicted combinations under some specified relationship through nearest-neighbor search. NNJE provides more powerful constraints for entity boundary recognition and entity pair combination, improves the model prediction accuracy and improves the model performance. The experiment set up a joint extraction data set of entity relationships in the process text. Experimental results show that NNJE improves the P score by 3.53% and the F1 score by 1.03% compared with the baseline model,and NNJE is better than PURE, CasRel, PRGC and TPlinker. It shows that the proposed method effectively improves the prediction results of entity-relation triples and improves model performance.

**Keywords:** process text , joint extraction , nearest neighbor

* 通讯作者
©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版
**基金项目:**辽宁省应用基础研究（2022JH2/101300248）国家自然科学基金（U1908216）

## 1 引言

实体关系联合抽取是自然语言处理、知识图谱构建的重要任务之一，其目的是从给定的非结构化文本中以（主体，关系，客体）的形式抽取实体对以及它们之间的关系。近年来，知识图谱在制造领域的广泛应用，使得工艺文本实体关系抽取任务尤为重要。

开展工艺文本实体关系联合抽取研究，相较于通用领域更为复杂。工艺文本中包含大量零件、材料特性、设备参数等专业领域知识信息，实体之间关系更为复杂。与通用领域人名、地名与机构名3类实体相比，工艺文本实体类粒度更细（18类），且工艺文本关系类数量远超于通用领域，如Tacrev数据集 (Alt et al., 2020)具有40种关系，SemEval (Hendrickx et al., 2010)数据集具有17种关系，而工艺文本具有62种关系，实体关系联合抽取难度更高。

实体关系联合抽取代表方法有DirectRel (Shang et al., 2022b)、TPLinker (Wang et al., 2020)与OneRel (Shang et al., 2022a)等。上述方法将实体关系联合抽取问题建模为二分图连接判断问题或实体边界字搭配判断问题。DirectRel模型把关系三元组抽取任务看做一个候选实体构成的二分图的连接问题，通过枚举所有小于定长的连续字符作为实体候选集合，该方式会产生过多的负样本，导致准确率较低。特别在工艺文本中实体的平均长度较通用领域（1.85个字符）更长，平均实体长度达3.85个字符（最长实体包含49个字符）。TPLinker与OneRel将三元组抽取问题建模为在特定关系下实体边界字搭配判断问题，这种填表方法改善了基于跨度的方法中实体抽取效果不佳的问题。但TPLinker只将关系类型视为可训练嵌入，实体和关系之间的相互约束不足，导致在组装三元组时效果不佳。OneRel将实体关系联合抽取作为细粒度三元组分类任务，相较于TPLinker采用更轻量级的填表方式。这种填表方式将实体边界进行搭配，但在工艺文本中具有实体边界字搭配的关系兼类歧义现象，导致OneRel抽取的实体关系三元组过度冗余，准确率较低。专业领域中实体关系联合抽取需要以该领域的知识为依据，并兼顾其语言规律特性，使得专业领域的实体关系联合抽取具有一定的难度。工艺文本实体间具有区别于通用领域的特殊字间搭配规律，在数据集规模小的情况下，势必会造成学习的困难。

为解决上述问题，本文提出一种最近邻联合抽取模型，用于工艺文本实体关系联合抽取任务。该方法利用工艺文本中实体边界字间搭配规律建模外显记忆，通过最近邻方法（Nearest Neighbor）为待预测组合检索出具有相似字间搭配的实例，将搭配错误的三元组进行删除以提高模型抽取准确率。实验结果表明，该文方法在工艺文本关系数据集下较基线模型OneRel的准确率P值提高了3.53%，F1值提升了1.03%，优于PURE (Zhong and Chen, 2021)、CasRel (Wei et al., 2020)、PRGC (Zheng et al., 2021)与TPlinker等方法。

KNN是一种基于实例学习的分类算法，适合多分类任务，不需要构建分类模型。有研究指出预训练语言模型在极低的资源条件下表现不稳定，因为在训练期间复杂的示例不容易被记忆到模型的嵌入（Embedding）向量中 (Chen et al., 2022)。目前，在机器翻译 (Zhu et al., 2023)，语言模型 (He et al., 2021)，关系抽取 (Wan et al., 2022)等任务中有加入了KNN外显记忆增强的方法，用来增强模型的鲁棒性与泛化能力。但在联合抽取，尤其是中文领域联合抽取还未见有所研究。

## 2 相关工作

实体关系联合抽取通常采用以下三种方法：基于参数共享的方法、基于联合解码的方法以及单模块单步骤方法。

### 2.1 基于参数共享的方法

参数共享的方法将实体关系联合抽取任务分解成实体识别和关系抽取两个子任务，子任务间共享编码层信息以增强交互。

Miwa等 (2016)提出基于端到端的树形结构联合抽取模型，首次采用神经网络的方法实现联合抽取。Zheng等 (2017a)提出了一种混合神经网络模型来提取实体及其语义关系，长短时记忆网络（Long Short Term Memory，LSTM）用于解码实体信息，卷积神经网络用于实现关系抽取。Bekoulis等 (2018)使用CRF层将实体识别任务和关系抽取任务建模为一个多头选择问题，关系抽取层对实体间每种关系做二分类预测。Yu (2020)等人提出了基于分解策略的联合抽取方法，先识别出所有的头部实体，然后对于每个主体识别出相关的客体及关系；最后把这两个任务转化成序列标注问题。Zeng等 (2018)提出了一种端到端神经模型，采用复制机制从句子中提取关系，解码器依次产生关系和主客体。Zeng等 (2020)提出了一个带有复制机


(c) 2024 中国中文信息学会计算语言学专业委员会

制的多任务学习框架。编码部分使用双向长短时记忆网络建模句子上下文信息，解码部分使用注意力机制与LSTM相结合建模，通过一个全连接层获取输出，结合复制机制生成多对三元组。Zheng等 (2021)为改进信息冗余和扩展性问题，提出一个新的端到端的框架PRGC,将实体关系联合抽取分解成了三个子任务：关系判断、实体抽取和头尾实体对齐，但该模型由于解码步骤复杂存在误差传递和暴露偏差问题。

### 2.2 基于联合解码的方法

基于联合解码的方法将实体识别和关系抽取两个任务映射在统一的框架结构下，进行全局的优化以及联合解码，加强模块之间的交互，实现实体关系的信息共享。Zheng等 (2017b)提出了一种基于序列标注的联合解码模型，在原有BIES标注方案上进行了扩展，使用BIEOS标签来指代实体的位置，并将关系类型融入标签中。Dai等 (2019)提出一种基于注意力机制的实体关系联合抽取模型，设计了一种可以同时表示实体类型和重叠关系的标记方案。Sui等 (2023)提出了一个非自回归解码的方法，使得各个三元组信息进行双向交互，并融合句子特征向量，用二分图匹配损失函数，针对无序集合的损失计算方式，从各种排列方式中找出和目标匹配损失最小的组合。

### 2.3 单模块单步骤方法

Shang等 (2022b)通过使用二分图实现对三元组的单步骤抽取。但该模型在数据处理时，生成候选实体集合的方式会产生过多的负样本，对实体平均长度较长的数据会使计算复杂性更高，导致准确率较低。Shang等 (2022a)提出直接从文本语句中抽取三元组，将联合抽取看作细粒度的三元组分类问题。但工艺文本实体和关系粒度较细，实体边界字搭配的关系兼类歧义现象多，导致OneRel抽取的实体关系三元组过度冗余，准确率较低。单模块单步骤模型通过全局矩阵或是图的方法实现从文本中直接抽取三元组，相较于参数共享和联合解码方法避免了级联错误的发生，针对工艺文本关系数据集中实体和关系较多，具有实体边界字搭配的关系兼类歧义现象的问题，单模块单步模型表现仍有不足。

## 3 最近邻联合抽取模型

### 3.1 问题定义

给定输入文本$S=\{x_1,x_2,x_3,\ldots,x_L\}$ 以及关系集合$R=\{r_1,r_2,r_3,\ldots,r_N\}$，实体关系联合抽取任务的目标是从给定文本中抽取出所有可能的三元组$T=\{(s_i,r_i,o_i)\}_{i=1}^{K}$，其中$L$为句子长度，$N$为关系数量，$K$为句子中三元组数量，$s$为主体实体，$o$为客体实体。

### 3.2 模型总框架

本文所提出的最近邻联合抽取模型NNJE（Nearest Neighbor Joint Extraction Model）沿袭了OneRel模型的标注方法：特定关系角标记法（Relation Specific Horns Tagging）。模型总框架如图 1所示，NNJE由OneRel和KNN模块组成，其中预训练语言模型BERT是Bert-base-chinese[0]模型。首先将序列$S=\{x_1,x_2,x_3,\ldots,x_L\}$输入到OneRel模型中，用BERT作为编码器捕捉每个字符的表征向量$\{\boldsymbol{e}_1,\boldsymbol{e}_2,\boldsymbol{e}_3,\ldots,\boldsymbol{e}_L\}$，并枚举出所有可能的$(\boldsymbol{e}_i,\boldsymbol{r}_n,\boldsymbol{e}_j)$组合。由两个全连接层构建出分数向量$\boldsymbol{v}_{(w_i,r_n,w_j)}$，将分数向量$\boldsymbol{v}$输入Softmax函数得到概率$P_{\text{OneRel}}\left(y_{(w_i,r_n,w_j)}|\boldsymbol{S}\right)$。同时将分数向量$v$作为查询向量输入到KNN模块中，KNN模块通过欧式距离函数$d(q,h)$计算出与待预测组合具有相似字间搭配规律的前$k$个键-值对，并计算概率$P_{\text{KNN}}\left(y_{(w_i,r_n,w_j)}|\boldsymbol{S}\right)$。将两个模块概率值加权求和得到最终概率$P_{\text{NNJE}}\left(y_{(w_i,r_n,w_j)}|\boldsymbol{S}\right)$，解码得到三元组抽取结果。

本章将分别介绍模型各模块的计算过程，包括角标记法、OneRel模型训练过程以及KNN外显记忆的构建与检索过程。

### 3.3 OneRel

OneRel将联合抽取任务转换成细粒度的三元组分类问题，用一种简洁的标注方法，特定关系角标记法，来标注三元组信息。这种角标记法从“BIE”(Begin,Inside,End)标注格式演化

[0]https://huggingface.co/bert-base-chinese


(c) 2024 中国中文信息学会计算语言学专业委员会

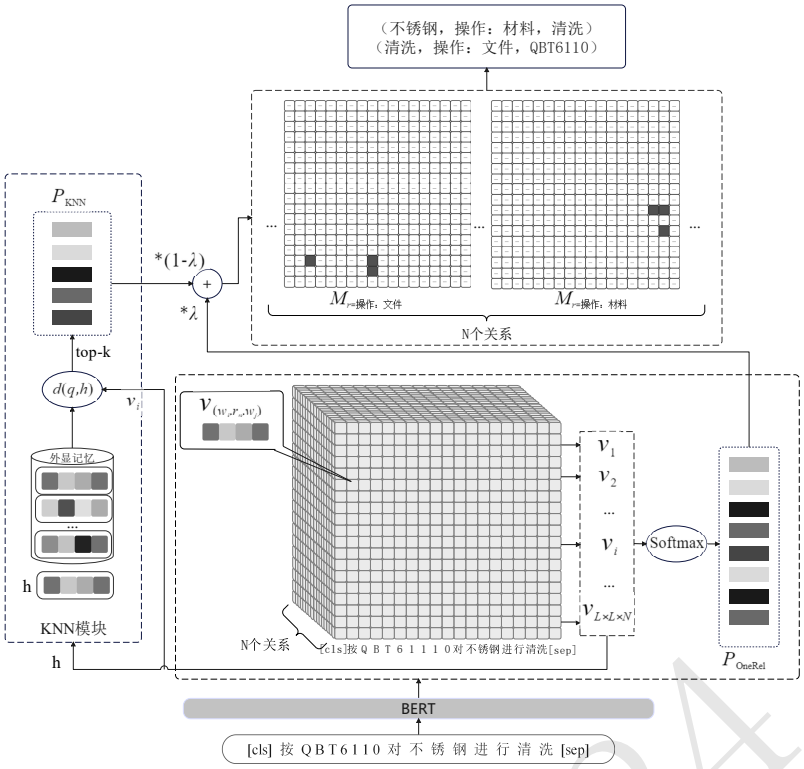



Figure 1: NNJE模型架构图

而来，并用“H”、“E”表示实体的开始以及结束。例如，HB指主体的开始位置，TE指客体的结束位置。OneRel把指定关系下的字间组合定义为四类，用于检测实体边界。(1)HB-TB：指在某种特定关系条件下主体和客体的起始位置标记。(2)HB-TE：指在某种特定关系条件下主体的起始位置和客体的结束位置标记。(3)HE-TE：指在某种特定关系条件下主体和客体的结束位置标记。(3)“-”：指在任何关系下都不构成实体边界的字符对，即除了上述情况下的所有字符对。如图 2所示，在句子“铆钉的铆接验收标准按QBT1111。”中有两个三元组，分别是：（铆钉，操作：文件，QBT1111）和（铆钉，操作：零件，铆接）。其中，在关系“操作：文件”下，主体和客体分别是“铆钉”和“QBT1111”。那么组合（铆，操作：文件，Q）就在子矩阵$\boldsymbol{M}_{r=\text{操作：文件}}$中被标记为“HB-TB”。同样的，组合（铆，操作：文件，1）被标记为“HB-TE”，组合（钉，操作：文件，1）被标记为“HE-TE”。不同关系将被标记在不同的子矩阵中。这种标记方法使得一个序列中所有的$(\boldsymbol{e}_i, \boldsymbol{r}_n, \boldsymbol{e}_j)$组合都能用一个三维张量$\boldsymbol{M}^{L\times N\times L}$表示。由于这种方法能够同时表示出主客体边界以及主客体之间的关系，因此解码时可以直接从$\boldsymbol{M}^{L\times N\times L}$中解码得到三元组。

OneRel模型首先将序列输入预训练语言模型BERT中进行编码，如式(1)。

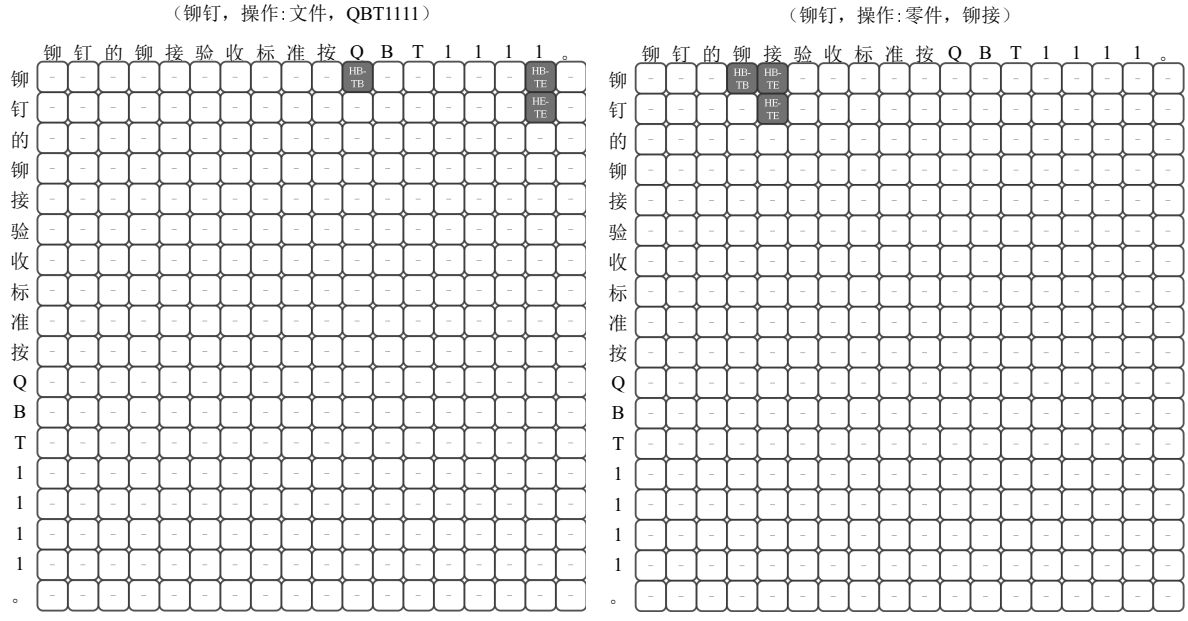
$$\{\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_L\} = BERT\left(\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_L\}\right) \tag{1}$$

然后枚举出所有可能的$(\boldsymbol{e}_i, \boldsymbol{r}_n, \boldsymbol{e}_j)$组合,通过一个分类器为每个组合分配高置信度标签。受知识图谱嵌入技术 (Nickel et al., 2016)的启发，该分类器使用两个全连接网络实现并行评分，可以自适应地学习从实体特征到实体对表示的映射函数，将分数函数定义为如式(2)。

$$\boldsymbol{v}_{(w_i,r_n,w_j)_{n=1}^{N}} = \boldsymbol{R}^T \phi(\text{drop}(\boldsymbol{W}[\boldsymbol{e}_i; \boldsymbol{e}_j]^T + \boldsymbol{b})) \tag{2}$$

其中，[;]表示拼接操作，$\boldsymbol{W} \in \mathbb{R}^{d_e \times 2d}$，$d_e$表示实体对表征的维度，$b$是可训练的权重。drop(·)是Dropout函数 (Srivastava et al., 2014)用来防止过拟合，$\phi(\cdot)$为RELU激活函数。$\boldsymbol{R} \in \mathbb{R}^{d_e \times 4N}$，其中4是角标记标签的数量，$\boldsymbol{R}$和$\boldsymbol{W}$均可以视为可训练的权重。$\boldsymbol{v}$是模型输出的分数向量。


(c) 2024 中国中文信息学会计算语言学专业委员会




(a) $M_{r=\text{操作：文件}}$角标记示例　　(b) $M_{r=\text{操作：零件}}$角标记示例

Figure 2: 角标记关系示例图

在OneRel模型预测时，将得到的分数向量输入到Softmax函数中，如式(3)，得到$(w_i, r_n, w_j)$组合的概率分布$P_{\text{OneRel}}(y_{(w_i,r_n,w_j)}|\boldsymbol{S})$。

$$P_{\text{OneRel}}(y_{(w_i,r_n,w_j)}|\boldsymbol{S}) = \text{Softmax}(\boldsymbol{v}_{(w_i,r_n,w_j)}) \tag{3}$$

损失函数定义如式(4)。

$$\mathcal{L}_{\text{triple}} = -\frac{1}{L \times N \times L} \times \sum_{i=1}^{L}\sum_{n=1}^{N}\sum_{j=1}^{L} \log P(y_{(w_i,r_n,w_j)} = t_{(w_i,r_n,w_j)}|\boldsymbol{S}) \tag{4}$$

其中，$t_{(w_i,r_n,w_j)}$为从数据集中获取的正确标签。

### 3.4 KNN

KNN的外显记忆数据是以键-值对的形式存在的，在KNN模块中，将训练数据中的序列输入到OneRel模块中，通过模型训练得出的每个$(w_i, r_n, w_j)$组合的分数向量$v_{(w_i,r_n,w_j)}$作为外显记忆集$D$中的键$h$。值$u$则是由每个$(w_i, r_n, w_j)$组合对应的角标记标签构成，如式(5)。

$$D = \{(\boldsymbol{h}, u)|(\boldsymbol{v}_{(w_i,r_n,w_j)}, t_{(w_i,r_n,w_j)})\} \tag{5}$$

KNN模块并不参与训练，仅利用实例之间的相似性进行预测完成分类任务。在模型预测阶段，将模型输出的分数向量$\boldsymbol{v}_{(w_i,r_n,w_j)}$作为KNN模块的输入，记作查询向量$v_i$，计算$v_i$与外显记忆集中每个键$h$的欧氏距离，取出与查询向量$\boldsymbol{v}_i$距离最小的$k$个键-值对，构成查询向量的最近邻集合$D^i$，以投票机制计算出当前查询向量的标签概率分布，如式(6)。

$$P_{\text{KNN}}(y_{(w_i,r_n,w_j)}|\boldsymbol{S}) \propto \sum_{(\boldsymbol{h},u)\in D^i} \text{I}_{y_i=u} \exp(\frac{-d(\boldsymbol{v}_i, \boldsymbol{h})}{T}), T \in R^+ \tag{6}$$

模型最终加权求和OneRel和KNN模型预测概率分布，如式(7)。

$$P_{\text{NNJE}}(y_{(w_i,r_n,w_j)}|\boldsymbol{S}) = (1-\lambda)P_{\text{OneRel}}(y_{(w_i,r_n,w_j)}|\boldsymbol{S}) + \lambda P_{\text{KNN}}(y_{(w_i,r_n,w_j)}|\boldsymbol{S}) \tag{7}$$

CCL 2024

## 4 实验数据与实验结果分析

### 4.1 实验数据

本文使用了人工标注的工艺文本关系数据集，数据来源为某型号飞机所遵照的工艺规范文件。实验数据集包含4421条句子，人工制定了62种关系类，共标注7646个关系，按照8:1:1的比例划分数据集为训练集、测试集和验证集。具体关系类及其数量分布情况见表 1。

| 关系子类 | 关系数量 | 关系子类 | 关系数量 | 关系子类 | 关系数量 |
|---|---|---|---|---|---|
| 操作:工艺辅料 | 830 | 属性：表注 | 165 | 工艺辅料：部位 | 21 |
| 操作：材料 | 111 | 属性：文件 | 101 | 材料：部位 | 18 |
| 操作：多余物 | 55 | 属性：图注 | 73 | 孔：部位 | 22 |
| 操作：零件 | 714 | 孔：表注 | 2 | 工具：部位 | 28 |
| 操作：部位 | 367 | 孔：文件 | 8 | 零件：部位 | 152 |
| 操作：孔 | 110 | 工具：图注 | 31 | 部位：属性值 | 34 |
| 零件：零件编号 | 156 | 工具：表注 | 25 | 零件：属性值 | 43 |
| 材料：材料牌号 | 23 | 工具：文件 | 8 | 工具：属性值 | 31 |
| 工具：工具编号 | 8 | 部位：文件 | 4 | 工艺辅料：属性值 | 10 |
| 工艺辅料：工艺辅料编号 | 771 | 工艺辅料：表注 | 22 | 操作：属性值 | 178 |
| 零件：属性 | 395 | 工艺辅料：文件 | 11 | 孔：属性值 | 9 |
| 操作：属性 | 454 | 零件：表注 | 6 | 多余物：属性值 | 6 |
| 工具：属性 | 31 | 零件：图注 | 17 | 材料：属性值 | 9 |
| 部位：属性 | 193 | 零件：文件 | 6 | 属性：属性值 | 796 |
| 工艺辅料：属性 | 169 | 部位：图注 | 6 | 数量：部位 | 11 |
| 孔：属性 | 94 | 操作：图注 | 75 | 数量：操作 | 16 |
| 材料：属性 | 46 | 操作：文件 | 376 | 数量：孔 | 16 |
| 部位：多余物 | 42 | 操作：表注 | 24 | 数量：零件 | 53 |
| 部位：工艺辅料 | 18 | 操作-用具 | 340 | 数量：工艺辅料 | 59 |
| 部位：零件 | 6 | 零件：材料 | 60 | 数量：工具 | 5 |
| 文件：表注 | 4 | 工具：材料 | 31 | | |

Table 1: 工艺文本数据集关系类型统计信息

### 4.2 对比方法

本文选择PURE(Zhong and Chen, 2021)、SpERT(Eberts and Ulges, 2020)、CasRel (Wei et al., 2020)、PRGC (Zheng et al., 2021)、TPLinker (Wang et al., 2020)和OneRel (Shang et al., 2022a)作为对比方法。其中，PURE将任务分成实体识别和关系判断两个子任务模型分别进行训练。SpERT、CasRel与PRGC是基于参数共享的实体关系联合抽取方法。SpERT先抽取出文本中的实体根据抽取出的实体对实体对做关系分类。CasRel采用层叠指针网络的思想，先识别所有可能的主体集合，在每个给定的关系类别下抽取与主体存在该关系的客体。PRGC采用端到端框架，先判断预测句子中可能存在的关系集合，再针对每个关系分别进行两次序列标注操作，分别提取出主体客体，再使用全局关联矩阵确定实体对。TPLinker是基于联合解码的方法，定义了3种链接方式判断实体边界，根据实体边界组合创建关系表格以抽取三元组。OneRel将实体关系联合抽取转换成细粒度的三元组分类问题，用特定关系角标注法标注实体边界，进行三元组抽取。

### 4.3 评价指标

本文使用准确率P（Precision）、召回率R（Recall）、F1值作为模型的评价指标，F1值越

CCL 2024

高代表抽取效果越好。评估方法由下式列出。

$$P = \frac{TP}{TP + FP} \tag{8}$$

$$R = \frac{TP}{TP + FN} \tag{9}$$

$$F = \frac{2 \cdot P \cdot R}{P + R} \tag{10}$$

其中，TP表示将正类抽取为正类的数量，指的是模型在实体关系联合抽取任务中，对输入实例的主体实体、客体实体以及实体对间的关系类型全部抽取正确的实例数量。相应地，FP为将负类抽取为正类数，FN为将正类抽取为负类数。

### 4.4 参数设置

本文实验采用了Torch 1.8.0框架。模型的超参数取值如表 2所示。在超参数中，k代表最近邻样本的数量，KNN模块的概率权重用 $\lambda$ 表示，$\lambda$ 的取值范围为[0.05,0.95]，本文选取该范围中在验证集上效果最好的 $\lambda$ 作为最终模型参数取值。

| 超参数 | 取值范围 |
|---|---|
| 训练轮次（Epoch） | 100 |
| 训练批大小（Batch Size） | 2 |
| 学习率（Learning Rate） | 5e-6 |
| k | 5 |
| $\lambda$ | [0.05,0.95] |

Table 2: 模型超参数设置

### 4.5 实验结果分析

本文选用当前效果较为突出的6个深度学习模型作为对比实验，在工艺文本关系数据集下进行评估，NNJE模型的F1值优于其他对比方法，结果如表 3所示。通过分析实验结果可进一步得出如下结论。

| 模型 | P | R | F1 |
|---|---|---|---|
| TPLinker | 37.53% | 24.42% | 29.58% |
| SpERT | 37.72% | 32.77% | 35.07% |
| PURE | 49.08% | 33.52% | 39.83% |
| CasRel | 50.39% | 34.18% | 40.73% |
| PRGC | 57.30% | **52.30%** | 54.70% |
| OneRel | 56.75% | 52.20% | 54.38% |
| NNJE | **60.28%** | 51.27% | **55.41%** |

Table 3: 实体关系联合抽取的实验效果

第二十三届中国计算语言学大会论文集，第406页-第417页，太原，中国，2024年7月25日至28日。
卷1：主会论文
(c) 2024 中国中文信息学会计算语言学专业委员会

| 模型 | 子任务 | P | R | F1 |
|---|---|---|---|---|
| OneRel | 实体识别 | 80.21% | 71.60% | 75.66% |
| | 实体对组成 | 58.57% | 53.76% | 56.06% |
| | 关系分类 | 57.62% | 53.00% | 55.22% |
| | 三元组 | 56.75% | 52.20% | 54.38% |
| NNJE | 实体识别 | **81.78%** | 69.90% | 75.37% |
| | 实体对组成 | **62.11%** | 52.51% | **56.91%** |
| | 关系分类 | **61.07%** | 51.94% | **56.13%** |
| | 三元组 | **60.28%** | 51.27% | **55.41%** |

Table 4: 实体关系联合抽取子任务实验效果

首先，NNJE模型F1值比次优模型OneRel高出1.03%，准确率P值提高了3.53%，验证了引入KNN方法能够有效提升单模块单步骤方法的模型性能。

其次，NNJE和OneRel模型抽取结果明显优于TPLinker、PURE和CasRel三个模型，说明在工艺文本实体关系联合抽取任务中，使用全局矩阵的方式优于其他各种分步骤分模块的抽取方式，这是因为全局矩阵能够充分捕捉工艺文本中实体与实体之间的特征关联，更适应工艺文本的复杂性以及专业性。与这两个模型相比，由于工艺文本关系数据集相较通用领域数据集粒度更细，而TPLinker的标注复杂度高，若关系数目大，解码部分矩阵参数量大，从而导致训练收敛速度慢和抽取性能不高等问题；SpERT在未标注数据中随机生成负样本，对数据质量的要求较高，因此在本文数据样本少且有大量混淆的情况下抽取效果差；PURE是两个模型独立训练，受级联错误影响，抽取效果不佳；CasRel在抽取客体时，需计算每个关系下对应的客体，容易导致关系冗余等问题。而本文模型NNJE优于OneRel与其他几种对比方法，验证了NNJE模型引入KNN方法的有效性。

此外，本文在不同子任务上进一步探索了NNJE模型效果，即实体识别、实体对组成以及关系分类任务。如表 4所示，在实体对组成和关系抽取子任务上，NNJE展现了优秀的性能。并且在实体识别F1值相当的情况下，NNJE模型的实体对组成、关系判断子任务的准确率和F1值都显著高于OneRel模型。这证明KNN通过查询与待预测组合有相似字间搭配规律的键-值对，增强了实体对识别以及关系分类性能。

### 4.6 KNN权值影响分析

本文探索了KNN权值$\lambda$对模型效果的影响，$\lambda$在[0.05,0.95]间取值，模型效果如图 3所示。从图中可以看到F1值随着$\lambda$的增长总体呈上升趋势，达到峰值后有所回落，除极端权值[0.05,0.10]和[0.85,0.95]外，当KNN权值越大时普遍效果更好。

### 4.7 实例分析

表 5中展示了文本、正确三元组、OneRel方法输出的三元组与NNJE方法输出的三元组，阐述了KNN模块对实体错误和关系分类错误两种错误三元组的删除情况。在表 5的实例1中，操作“压印”与部位“铣切边”和“凸缘边”之间不存在关系，部位“凸缘边”与属性“间距”应存在“部位：属性”的关系，。同样的，在实例2中OneRel模型错误地将铆钉头的零件编号“A-286”识别成工具“铣刀”的工具编号，把“硬质合金刀齿”错误识别成零件，并将不与任何实体有关系的数量实体“25个”和它们组成了三元组，KNN的加入删除了这些错误三元组。

在表 6中还展示了KNN对每个组合在正确标签上的预测概率及KNN检索出的实例等信息，从而进一步对KNN能够删除错误三元组的原因进行了分析。实例中，OneRel方法将零件实体“铆钉”和部位实体“周围的表面”识别成了“零件：部位”的关系。在工艺文本中“零件：部位”关系通常指零件操作时强调在某个部位操作，或零件以及零件本身的部位。在制造与装配工艺中，“钉”这类标准件不会对其“面”有特别操作与关注。在KNN的实例中，“钉”与“面”不会构成“零件：部位”关系下三元组中的实体边界，即主体尾和客体尾，故而将错误三元组删除。

从上述实例中可以看出，KNN能够根据实体边界的构成规律，构建实体边界字间的关系实例，通过检索拥有相似字间搭配规律的实例，有效地删除错误三元组，从而提高抽取准确率。


(c) 2024 中国中文信息学会计算语言学专业委员会

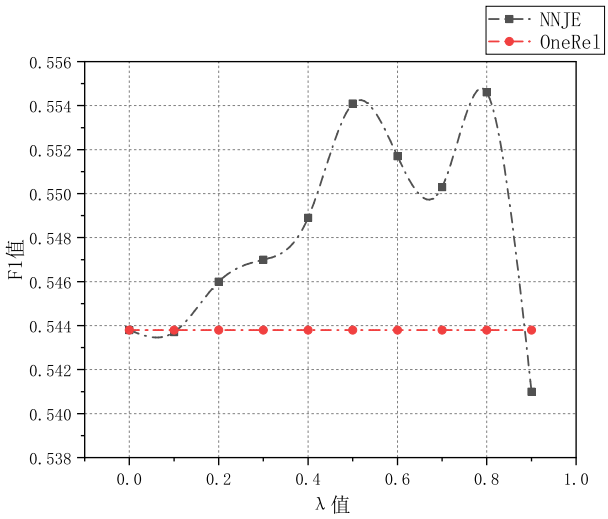



Figure 3: KNN权重变化下的F1值

| 实例1 | |
|---|---|
| 文本 | 在压印燃油传输槽口时，压印燃油传输槽口至零件边缘之间和压印燃油传输槽口至连接孔之间的最小间距如表6-3和图6-3所示，表6-3中的“E”尺寸适用于槽口至铣切边或凸缘边的间距。 |
| 正确三元组 | （压印，操作：部位，燃油传输槽口）（最小间距，属性:图注，表6-3）（压印，操作：部位，燃油传输槽口至连接孔之间）（“E”尺寸，属性:表注，表6-3）（最小间距，属性:图注，图6-3）（压印，操作：属性，最小间距）（压印，操作：部位，燃油，燃油传输槽口至零件边缘之间）（凸缘边，部位：属性，间距） |
| OneRel | （压印，操作：部位，凸缘边）（铣切边，部位：属性，最小间距）（凸缘边，部位：属性，最小间距）（最小间距，属性:图注，图6-3）**（压印，操作：部位，铣切边）（压印，操作：部位，燃油传输槽口）（最小间距，属性:图注，表6-3）** |
| NNJE | **（最小间距，属性:图注，表6-3）（最小间距，属性:图注，图6-3）（压印，操作：部位，燃油传输槽口）**（铣切边，属性:图注，最小间距） |
| 删除错误三元组 | （压印，操作：部位，凸缘边）（铣切边，部位：属性，最小间距）（凸缘边，部位：属性，最小间距）（最小间距，属性:图注，图6-3）（压印，操作：部位，铣切边）（压印，操作：部位，燃油传输槽口）（最小间距，属性:图注，表6-3） |
| **实例2** | |
| 文本 | 铣切A-286铆钉头时应采用ZT507B双向旋转的约有25个硬质合金刀齿的铣刀。 |
| 正确三元组 | （铣切，操作:零件，铆钉头）（铣刀，工具:工具编号，ZT507B）（铆钉头，零件:零件编号，A-286） |
| OneRel | （25个，数量:零件，硬质合金刀齿）**（铣切，操作:零件，硬质合金刀齿）**（25个，数量:零件，铆钉头）**（铣切，操作:零件，铆钉头）**（铣刀，工具:工具编号，A-286）（铆钉头，零件:零件编号，A-286） |
| NNJE | **（铣切，操作:零件，铆钉头）（铆钉头，零件:零件编号，A-286）** |
| 删除错误三元组 | （25个，数量:零件，硬质合金刀齿）（铣切，操作:零件，硬质合金刀齿）（25个，数量:零件，铆钉头）（铣刀，工具:工具编号，A-286） |

Table 5: 删除错误三元组实例分析

| 文本 | 按NACA铆接方法正确安装的铆钉见图6-3，注意铆钉周围的表面要光滑。 | | | | |
|---|---|---|---|---|---|
| 正确三元组 | （铆钉，零件:图注，图6-3） | | | | |
| OneRel | （安装，操作:零件，铆钉）**（铆钉，零件:图注，图6-3）**<br>（铆钉，零件:部位，周围的表面） | | | | |
| NNJE | （安装，操作:零件，铆钉）**（铆钉，零件:图注，图6-3）** | | | | |
| （字符，关系，字符，标签）组合 | $P_{\text{OneRel}}$ | $P_{\text{NNJE}}$ | 检索出组合及标签 | | 检索出组合所在句子 |
| | | | 组合 | 标签 | |
| （“钉”，“零件:部位”，“面”，“-”） | 0.402 | 0.56 | （“纸”，“工艺辅料：工艺辅料编号”，“L”） | - | 如果表面用QBT6520揩布 或QBT640中性牛皮**纸**进行了保护，则允许在24小时内完成硅底胶或CMS-S**L**系列胶粘剂的施工。 |
| | | | （“条”，“工艺辅料：工艺辅料编号”，“7”） | HE-TE | 用真空袋密封胶**条**（QBT5537、QBT5537或QBT5537）将真空袋（QBT5664、QBT5664或QBT5664）密封到模具上，确保真空袋不架桥。 |
| | | | （“胶”，“工艺辅料：工艺辅料编号”，“1”） | - | 如有凸起或阶差允许 用QBT923（150目或更细）砂纸打磨修平，拼接缝内若有凹坑，允许使用CMS-AD-106发泡**胶**进行修补，修补后拼接缝及泡沫表面应用可剥布（QBT6000，QBT600**1**）或脱模布（QBT3463）铺贴覆盖。 |
| | | | （“件”，“工艺辅料：工艺辅料编号”，“理”） | - | 应力消除：通常，工程图样规定焊后进行热处**理**的低合金钢和碳钢零件的闪光焊缝以及不锈钢零**件**的闪光焊缝不要求消除应力。但是HY-Tuf钢（高韧钢），4140，4340和300M低合金钢零件则要求在焊后冷却至93℃～204℃(200～400)温度范围后进行消除应力，其消除应力的方法，可以用闪光焊机来加热焊缝或用感应线圈将其加热到538℃～649℃(1000～1200)的温度范围，也可以按QBT5000消除应力。 |

Table 6: KNN检索实例分析

CCL 2024

## 5 总结

本文利用工艺文本中实体边界字间搭配规律，提出面向工艺文本实体与关系的最近邻联合抽取模型NNJE。实验结果表明：首先，本文提出的最近邻联合抽取模型NNJE效果优于对比方法，表明了NNJE模型的优越性；其次，从NNJE模型与单模块单步骤模型OneRel的对比结果看，KNN能够有效增强抽取效果，且当KNN权值越高时，效果越好；再次，分别在各子任务上研究了NNJE模型，在实体对识别以及关系分类子任务上均有着较好的抽取效果；最后，通过实例分析，NNIE模型能够根据实体边界的构成规律，构建实体边界字间的关系实例，为待预测组合检索出与之字间搭配相似的键-值对。未来工作中，将对填表方式进行压缩，减少由于关系类数量过多导致产生过多负样本，构建更适用于多关系类别数据的抽取模型。

## 参考文献

Christoph Alt, Aleksandra Gabryszak, and Leonhard Hennig. 2020. TACRED revisited: A thorough evaluation of the TACRED relation extraction task. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1558–1569. Association for Computational Linguistics.

Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. 2018. Joint entity recognition and relation extraction as a multi-head selection problem. *Expert Systems with Applications*, 114:34–45.

Xiang Chen, Lei Li, Ningyu Zhang, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022. Relation extraction as open-book examination: Retrieval-enhanced prompt tuning. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2443–2448.

Dai Dai, Xinyan Xiao, Yajuan Lyu, Shan Dou, Qiaoqiao She, and Haifeng Wang. 2019. Joint extraction of entities and overlapping relations using position-attentive sequence labeling. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6300–6308.

Markus Eberts and Adrian Ulges. 2020. Span-based joint entity and relation extraction with transformer pre-training. In *ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020)*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 2006–2013.

Junxian He, Graham Neubig, and Taylor Berg-Kirkpatrick. 2021. Efficient nearest neighbor language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 5703–5714. Association for Computational Linguistics.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval@ACL 2010, Uppsala University, Uppsala, Sweden, July 15-16, 2010*, pages 33–38. The Association for Computer Linguistics.

Makoto Miwa and Mohit Bansal. 2016. End-to-end relation extraction using lstms on sequences and tree structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Maximilian Nickel, Lorenzo Rosasco, and Tomaso Poggio. 2016. Holographic embeddings of knowledge graphs. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

Yu-Ming Shang, Heyan Huang, and Xianling Mao. 2022a. Onerel: Joint entity and relation extraction with one module in one step. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 11285–11293.

Yuming Shang, Heyan Huang, Xin Sun, Wei Wei, and Xian-Ling Mao. 2022b. Relational triple extraction: One step is enough. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 4360–4366. ijcai.org.


(c) 2024 中国中文信息学会计算语言学专业委员会

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Dianbo Sui, Xiangrong Zeng, Yubo Chen, Kang Liu, and Jun Zhao. 2023. Joint entity and relation extraction with set prediction networks. *IEEE Transactions on Neural Networks and Learning Systems*.

Zhen Wan, Qianying Liu, Zhuoyuan Mao, Fei Cheng, Sadao Kurohashi, and Jiwei Li. 2022. Rescue implicit and long-tail cases: Nearest neighbor relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 1731–1738. Association for Computational Linguistics.

Yucheng Wang, Bowen Yu, Yueyang Zhang, Tingwen Liu, Hongsong Zhu, and Limin Sun. 2020. Tplinker: Single-stage joint extraction of entities and relations through token pair linking. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 1572–1582. International Committee on Computational Linguistics.

Zhepei Wei, Jianlin Su, Yue Wang, Yuan Tian, and Yi Chang. 2020. A novel cascade binary tagging framework for relational triple extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1476–1488. Association for Computational Linguistics.

Bowen Yu, Zhenyu Zhang, Xiaobo Shu, Tingwen Liu, Yubin Wang, Bin Wang, and Sujian Li. 2020. Joint extraction of entities and relations based on a novel decomposition strategy. In *ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August-8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020)*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pages 2282–2289. IOS Press.

Xiangrong Zeng, Daojian Zeng, Shizhu He, Kang Liu, and Jun Zhao. 2018. Extracting relational facts by an end-to-end neural model with copy mechanism. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 506–514.

Daojian Zeng, Haoran Zhang, and Qianying Liu. 2020. Copymtl: Copy mechanism for joint extraction of entities and relations with multi-task learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9507–9514.

Suncong Zheng, Yuexing Hao, Dongyuan Lu, Hongyun Bao, Jiaming Xu, Hongwei Hao, and Bo Xu. 2017a. Joint entity and relation extraction based on a hybrid neural network. *Neurocomputing*, 257:59–66.

Suncong Zheng, Feng Wang, Hongyun Bao, Yuexing Hao, Peng Zhou, and Bo Xu. 2017b. Joint extraction of entities and relations based on a novel tagging scheme. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1227–1236. Association for Computational Linguistics.

Hengyi Zheng, Rui Wen, Xi Chen, Yifan Yang, Yunyan Zhang, Ziheng Zhang, Ningyu Zhang, Bin Qin, Xu Ming, and Yefeng Zheng. 2021. PRGC: potential relation and global correspondence based joint relational triple extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6225–6235. Association for Computational Linguistics.

Zexuan Zhong and Danqi Chen. 2021. A frustratingly easy approach for entity and relation extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 50–61. Association for Computational Linguistics.

Wenhao Zhu, Jingjing Xu, Shujian Huang, Lingpeng Kong, and Jiajun Chen. 2023. INK: injecting knn knowledge in nearest neighbor machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 15948–15959. Association for Computational Linguistics.


(c) 2024 中国中文信息学会计算语言学专业委员会

# 融合扩展语义和标签层次信息的文档级事件抽取

**符玉娇[1]，廖健[1]，李旸[2]，郭张峰[1]，王素格[1,3,*]**
1.山西大学计算机与信息技术学院，山西省 太原 030006
2.山西财经大学金融学院，山西省 太原 030006
3.山西大学计算智能与中文信息处理教育部重点实验室，山西省 太原 030006
wsg@sxu.edu.cn

## 摘要

文档级事件抽取是自然语言处理中的一项重要任务，面临论元分散和多事件提及的挑战，现有研究通常从文档的所有句子中抽取论元，通过论元角色建模捕获实体间关系，忽略了文档中事件-句子间的关联差异性。本文提出了一种融合扩展语义和标签层次信息的文档级事件抽取方法。首先，利用大语言模型对文本和事件类型标签与论元角色标签进行语义扩展，以引入更丰富的背景语义信息；其次，基于关联差异性的事件类型检测模块，获取文档中与事件类型高度相关的句子，通过约束候选实体的抽取范围，来缓解论元分散问题；进一步，针对文档提及的多个事件类型，利用有向无环图从候选实体中抽取论元，获取所有事件要素。在ChFinAnn和DuEE-Fin两个数据集上的实验结果表明，本文提出的方法相比基线模型可以有针对性地缓解多个事件所属论元分散的问题，有效地提升事件抽取的性能。

**关键词：** 文档级事件抽取；标签语义扩展；关联差异建模；论元抽取

# Document-Level Event Extraction with Integrating Extended Semantics and Label Hierarchy Information

**Yujiao Fu[1]，Jian Liao[1]，Yang Li[2]，Zhangfeng Guo[1]，Suge Wang[1,3,*]**
1.School of Computer and Information Technology,
Shanxi University,Shanxi 030006
2.Shanxi University of Finance and Economics,Shanxi 030006
3.Key Laboratory Computational Intelligence and Chinese Information
Processing of Ministry of Education, Shanxi University,Shanxi 030006
wsg@sxu.edu.cn

## Abstract

Document-level event extraction is an important task in natural language processing, facing the challenges of scattered arguments and multi event mention, existing research typically extracts arguments from all sentences in a document and captures relationships between entities through argument role modeling, thereby ignoring the differences in association between events and sentences in documents. This paper proposes a document-level event extraction method that integrates extended semantic and hierarchical label information. Firstly, using large language model to semantically extend text and event type labels along with role labels to introduce richer contextual semantic information. Secondly, based on the event type detection module of correlation difference, sentences highly related to the event type in the document are obtained, and the problem of argument dispersion is alleviated by constraining the extraction range of candidate entities. Furthermore, for the multiple event types mentioned in the document, arguments are extracted from candidate entities using directed acyclic graphs to obtain all event elements. Experimental results on the ChFinAnn and DuEE-Fin datasets indicate that the proposed method effectively alleviates the problem of


(c) 2024 中国中文信息学会计算语言学专业委员会

scattered arguments belonging to multiple events compared to baseline models, and effectively improve the performance of event extraction.

**Keywords:** Document-level event extraction , Label semantic extension , Differences in association modeling , Argument extraction

## 1 引言

事件抽取是从文本中抽取有关特定事件的信息，即识别事件的类型和抽取事件的核心要素，包括执行动作的主体、客体、时间、地点等(Li et al., 2022)，进而将非结构化事件文本转化为结构化信息，这些信息可服务于事理图谱构建(Du et al., 2022)、基于事件的推荐系统(李治, 2022)等下游任务。针对事件相关要素分散于多个句子的问题，现有的句子级的文本事件抽取方法(Sheng et al., 2021; Lu et al., 2021)无法全面、完整地抽取出事件信息。因此，文档级事件抽取受到越来越广泛的关注，通过从长文本中抽取描述特定事件的关键信息，利用论元集合表示不同的事件(Liang et al., 2022)，可以获得更全面、细致的事件结构化信息，图1展示了一个包含3个事件的文档示例。

……

[S3]江西国泰集团股份有限公司（以下简称“公司”）收到公司持股5%以上股东江西鑫安信和投资集团有限责任公司（以下简称“鑫安信和”）部分股份解除质押的口头通知，具体情况如下：一、本次股份解除质押的具体情况

[S4]鑫安信和于2017年9月29日将其持有的公司27,020,000股无限售流通股（占公司总股本的比例为6.91%），质押给华泰证券股份有限公司（以下简称“华泰证券”），办理股票质押式回购交易业务（详见公司2017临065号公告）

[S5]2018年9月25日鑫安信和提前购回部分本金，同时对上述流通股质押办理延期（详见公司2018临087号公告）。2019年5月27日，鑫安信和将其中11,046,000股解除质押，并办理完成了相关手续。本次解除质押股数占公司总股本的2.82%

[S6]2019年7月26日，鑫安信和将其中7,364,000股解除质押，并办理完成了相关手续，本次解除质押股数占公司总股本的1.88%。二、持股5%以上股东股权质押情况 截至本公告披露之日，鑫安信和持有公司股份40,600,000股，占公司总股本的10.38%

[S7]已累计质押公司股份12,510,000股，占其持有公司股份总数的30.81%，占公司总股本的3.20%。特此公告

江西国泰集团股份有限公司 董事会 二〇一九年七月三十日

| 事件 | 事件类型 | 质押物所属公司 | 质押方 | 事件时间 | 质押股票/股份数量 | 质押物占总股比 | 质押物 | 披露时间 |
|---|---|---|---|---|---|---|---|---|
| 事件1 | 解除质押 | 江西国泰集团股份有限公司 | 江西鑫安信和投资集团有限责任公司 | 2019年5月27日 | 11,046,000 | 2.82% | 股份 | 二〇一九年七月三十日 |
| 事件2 | | | | 2019年7月26日 | 7,364,000 | 1.88% | | |
| 事件3 | 质押 | | | 2017年9月29日 | 27,020,000 | NULL | 无限售流通股 | |

图 1: 包含3个事件的文档示例，[•]为句子编号

对于文档级事件抽取，将面临以下两个挑战:

（1）论元分散，事件论元分散于一篇文章的多个句子中。例如图1，事件类型为“解除质押”的事件1中，句子[S3]的“江西国泰集团股份有限公司”和“江西鑫安信和投资集团有限责任公司”扮演“质押物所属公司”和“质押方”的角色，句子[S5]的“11046000”和“2.82%”扮演“质押股票/股份数量”和“质押物占总股比”的角色，句子[S7]的“二零一九年七月三十日”扮演“披露时间”的角色。针对论元分散问题，需要捕捉关联性较强的上下文来抽取完整的事件信息。

（2）多事件提及，即一篇文章包含多个事件，这些事件可能属于同一种事件类型，也可能属于不同的事件类型。例如图1，事件1和3的事件类型分别为“质押”和“解除质押”。多事件提及意味着确定文档包含的事件个数较为困难，因此需要通过事件-句子间的关联差异性约束候选实体的抽取范围。

针对论元分散问题，现有研究(Liang et al., 2022; Wan et al., 2023)主要利用注意力机制或图模型建立论元之间的关系；针对多事件提及问题，现有方法(Xu et al., 2021; Wang et al.,

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版
* 通讯作者 Corresponding Author.
**基金项目：**国家重点研发计划课题(2022QY0300-01)；国家自然科学基金项目(62106130,62376143,62076158)；山西省基础研究计划项目(202303021211021)；CCF-智谱AI大模型基金(CCF-Zhipu202310)

2023)利用跟踪机制或建立伪事件节点之间的联系，捕获事件间的依赖。在这些事件论元抽取研究中，针对同一事件类型，并未考虑句子与论元、事件类型间的关联差异性。而每篇文章中的每个句子与特定事件类型的相关性是有差异的，事件论元通常分散在与事件类型相关性较高的句子中，因此，需要建立事件论元与事件类型的关系。

本文提出了融合扩展语义与标签层次信息的文档级事件抽取模型(Document-Level Event Extraction with Integrating Extended Semantics and Label Hierarchy Information，IESLHI)。首先，利用大语言模型丰富的背景语义知识，构建增强提示，对文章和事件类型标签以及论元角色标签进行语义扩展，更好地理解和表示事件文本。在此基础上，对论元和事件类型的相关性进行建模，构建基于关联差异性的事件类型检测模块，获得文章中与每个事件类型标签相关性较高的句子，并基于该相关性对每个句子中抽取的实体给予不同程度的关注，建立候选实体。进一步，基于候选实体构建有向无环图，对文档包含的每个事件类型分别通过路径扩展进行论元抽取，以获得各事件的所有论元。通过在两个公开数据集上进行实验，实验结果验证了本文所提方法的有效性。

综上所述，本文的贡献如下：

(1)提出了一个融合扩展语义与标签层次信息的文档级事件抽取模型(IESLHI)，从文档中捕获与事件类型标签关联较强的句子。在此基础上，通过在论元抽取过程中聚焦强关联的句子，以约束候选实体的抽取范围，并针对每个事件类型构建向无环图抽取论元，提升文档级事件抽取的性能。

(2)提出了一种基于大语言模型的层次化标签语义扩展与融合方法，丰富文档内容和事件类型标签与论元角色标签的语义，有效地增强了模型理解文档信息和层次信息的能力。

(3)在两个公开数据集ChFinAnn和DuEE-Fin上进行实验，与已有的文档级事件抽取方法相比，所提模型的实验性能具有一定的提升。同时与开源大语言模型进行对比，所提模型在文档级事件抽取性能上优于大语言模型。本文的实现代码在论文发表后可在GitHub上获取。

## 2 相关工作

### 2.1 给定事件类型的论元抽取

论元抽取是在给定事件类型的情况下，抽取与事件有关的实体。之前的研究(Li et al., 2021; Ma et al., 2022) 通过构建特定事件类型的提示模版直接抽取论元,无需进行命名实体识别。但是，文档中的冗余信息导致抽取论元的性能较差。为了减少冗余信息，Pouran Ben Veyseh等人(2022)基于句法结构的最优传输排除不相关单词，并利用正则化对不相关单词的贡献进行约束；Xu等人(2022)引入抽象意义表示图构建句内和句间关系，并利用边界损失增强论元的边界判断。此外，为了融合丰富的语义信息，Liu等人(2023b)利用注意力机制融合触发词与上下文信息以及论元角色间的关系，进而预测跨度的角色。为了增强论元抽取过程中的可解释性，Liu等人(2023a)利用链式推理范式进行论元抽取，通过推理链的相互依赖性捕获句子间的长距离依赖；Ren等人(2023)引入基于相似性的检索增强策略给每个实例采样伪示例，提高模型的类比能力。

上述研究主要利用句间信息和实体间信息提高论元抽取的性能，忽略了事件类型和论元角色间的层次信息，本文旨在利用大语言模型对层次标签进行语义扩展，丰富事件类型标签与论元角色标签的语义。

### 2.2 文档级事件抽取

由于给定事件类型的论元抽取存在现实需求的局限性，因此，文档级事件抽取广受关注。Zheng等人(2019)基于有向无环图的路径扩展方法抽取论元，但它需要预先定义论元角色的抽取顺序。为了避免预先定义论元角色的顺序，Huang等人(2023)采用预填充策略，并行地生成事件记录的论元。Liang等人(2022)利用关系增强的注意力机制捕获实体间关系，但它未捕获句间信息，而篇章中的事件论元可能分散在多个句子中，捕捉跨句的实体信息对于文档级事件抽取非常重要。因此，许多研究利用图模型建立实体间关系以抽取论元，Zhu等人(2022)建立实体提及与伪触发词的单向边以及伪触发词间的双向边构造修剪完全图，利用伪触发词信息提升文档级事件抽取的性能。Wan等人(2023)建立token与token之间的双向边，通过训练集构造邻接矩阵，预测token之间边的关系。但是，这些方法忽略了文档中多事件间的相互依赖，Xu等人(2021)构建句子和实体提及的异质图捕获实体与句子之间的关系，同时，引入跟踪机制捕获


(c) 2024 中国中文信息学会计算语言学专业委员会

事件之间的相互依赖。Wang等人(2023)建立实体提及、句子与伪事件节点之间的关系，通过伪事件节点间的相互联系从全局上捕获事件信息。

上述研究旨在捕获实体与所有句子间的关系以解决论元分散问题，未考虑事件类型与句子的相关性问题，本文通过研究事件类型与句子之间的关联差异性来缓解论元分散。

## 3 文档级事件抽取模型

### 3.1 模型概述

针对第1节引言中提出的挑战，本文提出了一个融合扩展语义与标签层次信息的文档级事件抽取模型(IESLHI)，整体框架如下图2所示，包括三个核心部分：基于大语言模型语义扩展的实体抽取与表示、基于关联差异性的事件类型检测，和基于有向无环图的候选实体的论元抽取。首先，利用大语言模型解释文章的每个句子，得到语义扩展文本。再将事件类型标签作为查询来检测文章包含的事件类型。其次，对于文章包含的每个事件类型，判断每个句子与特定事件类型的相关性。最后，基于实体的有向无环图，对每个事件类型相关性较高的句子的实体进行路径扩展，迭代地生成每个事件的论元。

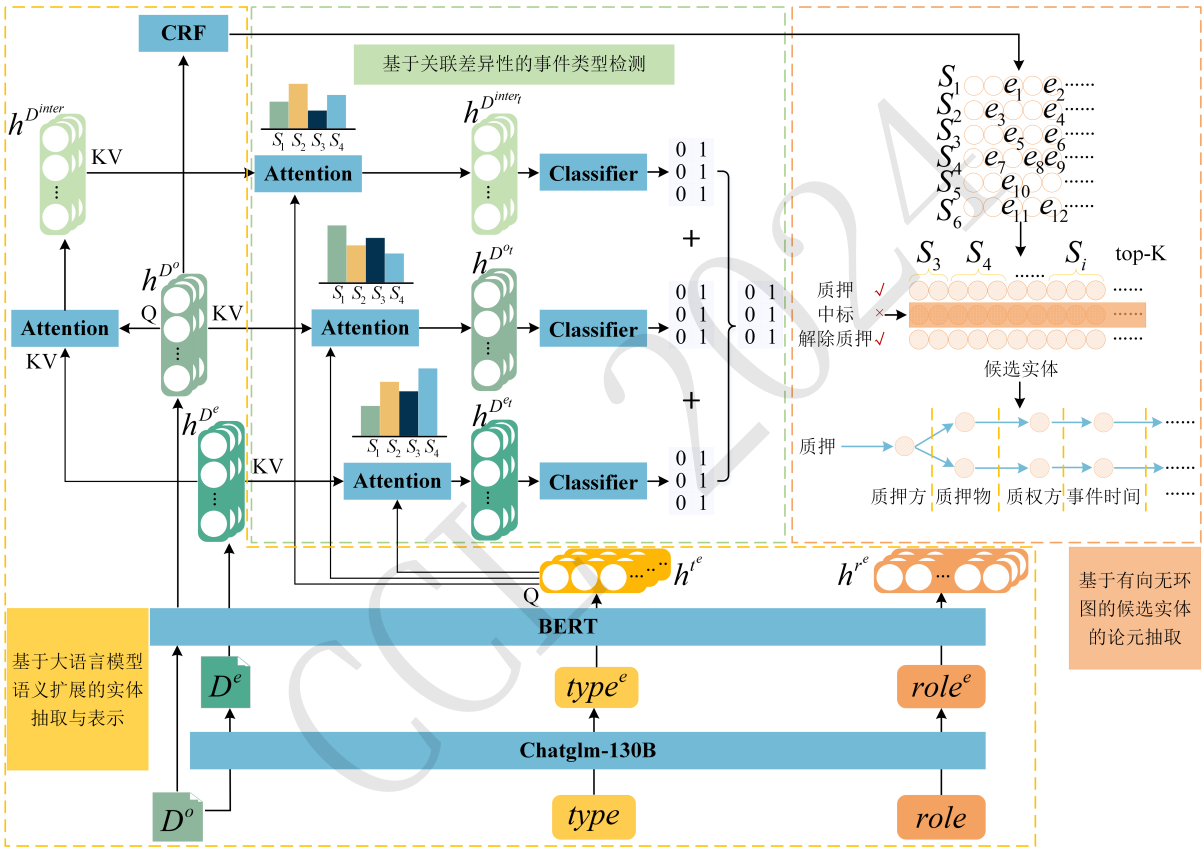



图 2: IESLHI模型框架

### 3.2 基于大语言模型语义扩展的实体抽取与表示

#### 3.2.1 实体抽取与表示

文档$D_i$由句子$S_1, S_2, ..., S_n$组成，$n$为文档$D_i$的句子个数。其中，每个句子由$m$个token组成，表示为$S_i = [W_{i1}, W_{i2}, ..., W_{im}]$，$W_{ij}$表示第$i$个句子中的第$j$个token，$W_{ij} \in R^d$，$d$为嵌入表示维度，$S_i \in R^{d \times m}$，$D_i \in R^{d \times m \times n}$。

为确定文档中的实体提及，需要对文档D中的句子$S_i$包含的命名实体进行识别。这里采用BERT模型，按句对token序列进行编码，得到包含句内上下文信息的表示$h_{ij}^W$。

$$h_{ij}^W = \text{BERT}(W_{ij}), h_i \in R^{d \times m} \quad (1)$$


(c) 2024 中国中文信息学会计算语言学专业委员会

句中实体标签按照BIOSE(Begin, Inside, Outside, Single, End)进行序列标注。将每一句中的token表示输入到CRF层得到标签序列的预测概率$p^{ner}$。

$$p^{ner} = \mathrm{CRF}(h_i) \tag{2}$$

实体提及为文档中处于不同位置的同一实体，且实体提及通常包含多个token。因此，对每个实体提及进行max pooling操作得到相应的嵌入表示。最后，对多个实体提及进行max pooling操作得到每个实体的表示，该表示服务于后续的论元抽取任务。

命名实体识别部分的损失函数为：

$$L_{ner} = -\sum_{i=1}^{n}\sum_{j=1}^{m}\log p_{ij}^{ner} \tag{3}$$

其中，$p_{ij}^{ner}$表示文档D中第$i$个句子的第$j$个token对应的预测概率。

### 3.2.2 文本语义扩展

在文档中，承接上下文的句子通常省略主体信息、句子的表达存在多义性。通过对文本的语义进行拓展，可以深入理解文本的逻辑关系和含义。本文利用大语言模型的提示方法丰富和补充每个句子的语义，保持句子的主体不变性以避免生成解释过程中引起的语义偏差。本文使用glm-3-turbo分析文本中词语和短语之间的语义相关性，构造提示模版Prompt1=“请结合上下文解释以下这句话：[句子]”解释每个句子，挖掘句子之间的语义相关性，生成感知上下文的语义扩展文本$D^e$。

对于得到的语义扩展文本，同样使用BERT对解释文档$D^e$的token序列进行编码，得到语义扩展文本对应的token表示$h_{ij}^e$，随后通过max pooling操作得到相应的句子嵌入表示$h_i^e$。

$$h_{ij}^e = \mathrm{BERT}(W_{ij}^e) \tag{4}$$

$$h_i^e = \mathrm{MaxPooling}(h_{i1}^e, ..., h_{im}^e), h_i^e \in R^d \tag{5}$$

其中，$h_{ij}^e$为解释文档$D^e$中第$i$个句子的第$j$个token的表示，$h_i^e$为解释文档$D^e$中第$i$个句子的表示，每个句子由$m$个token组成，$d$为嵌入表示维度。

### 3.2.3 标签层次信息解释

事件类型反映了一组事件的特性，但是，简短的事件类型名称具有模糊性，仅通过事件类型名称不能反映事件类型本身的丰富信息。利用大语言模型加以提示，构造提示模版Prompt2=“金融领域，怎样定义[事件类型]”解释每个事件类型标签，有助于理解文本描述的事件，得到具有丰富语义的、可以充分反映事件类型信息的事件类型解释文本，进而生成感知事件类型的解释文本表示。此外，每个事件类型定义了一系列论元角色，论元角色表示动词或其他词涉及到的实体扮演的角色，描述了实体之间的关系。但是角色名称本身表达的语义有限，利用大语言模型对角色进行详细的描述与扩展，构造提示模版Prompt3=“金融领域的[事件类型]事件中怎样定义[论元角色]”，得到有助于理解文本语义信息的论元角色解释文本，进而生成感知论元角色的解释文本表示。

与文本语义扩展类似，此处使用BERT对事件类型解释文本$t^e$、论元角色解释文本$r^e$进行编码，得到感知事件类型的解释文本表示$h^{t^e}$，感知论元角色的解释文本表示$h^{r^e}$，分别将其作为事件类型和论元角色的表示，$h^{t^e} \in R^d$，$h^{r^e} \in R^d$。

## 3.3 基于关联差异性的事件类型检测

对于文档D，通过注意力机制对各句子[CLS]表示$S_i^C$进行加权，得到不同事件类型下相应的文档表示$D^{o_t}$。其中，注意力的查询项为事件类型$h^{t^e}$。对于每一个事件类型，使用神经网络分类层判断其是否在文档中被提及，并得到相应事件类型的概率$p^{D^{o_t}}$，具体操作如下：

$$\alpha_i = \mathrm{softmax}(\frac{h^{t^e} \cdot S_i^C}{\sqrt{d_k}}) \tag{6}$$


(c) 2024 中国中文信息学会计算语言学专业委员会

$$D^{o_t} = \sum_{i=1}^{n} \alpha_i S_i^C \tag{7}$$

$$p^{D^{o_t}} = \text{softmax}(\text{FC}(D^{o_t})) \tag{8}$$

同理，对于解释文档$D^e$，使用注意力机制得到不同事件类型下的文档表示$D^{e_t}$。用神经网络分类层来判断每个事件类型是否被识别，并得到相应事件类型的概率$p^{D^{e_t}}$。

$$p^{D^{e_t}} = \text{softmax}(\text{FC}(D^{e_t})) \tag{9}$$

进一步，为将原始文档与语义扩展文本进行交互，得到融合上下文解释信息的文本表示，本模型将原始文档表示$[S_1, S_2, ..., S_n]$与语义扩展文本$[S_1^e, S_2^e, ..., S_n^e]$送入Transformer，得到其交互表示$[S_1^{inter}, S_2^{inter}, ..., S_n^{inter}]$。

$$S_i^{inter} = \text{Transformer}(Q, K, V) = \text{Transformer}(S_i^e, S_i, S_i) \tag{10}$$

通过注意力得到相应的文档表示$D^{inter}$，并完成事件类型检测，得到相应事件类型的概率$p^{D^{inter_t}}$。

$$p^{D^{inter_t}} = \text{softmax}(\text{FC}(D^{inter_t})) \tag{11}$$

将三种方式得到的概率求和，并进行归一化操作得到最终的相应事件类型的概率$p^{D^t}$。

$$p^{D^t} = \text{softmax}(p^{D^{o_t}} + p^{D^{e_t}} + p^{D^{inter_t}}) \tag{12}$$

事件类型检测的损失函数为：

$$L_{dect} = -\sum_{t=1}^{z} (\log p^{D^t} + \log p^{D^{o_t}} + \log p^{D^{e_t}} + \log p^{D^{inter_t}}) \tag{13}$$

其中，$z$表示事件类型的个数。

### 3.4 基于有向无环图的候选实体的论元抽取

由于事件论元抽取是为文档所包含事件的每个论元角色填充相应的实体，本文受(Zheng et al., 2019)的启发，使用基于实体的有向无环图生成每个事件的论元。预先定义每种事件类型的论元角色顺序，每个事件记录的长度等于预先定义的论元角色的数量，每个论元节点用实体或空节点填充。当文档D包含特定的事件类型时，按照论元角色顺序预测每个角色的实体。对于每个论元角色，可以存在0个、一个或几个满足条件的实体，因此每个论元角色展开的分支数量是不确定的。每一条完整路径表示预测出当前事件类型的一个事件。

针对事件类型与句子间的关联差异性，通过判断特定事件类型下相关性较高的句子，有助于检索和查找与该事件类型相关的事件记录，这些句子往往也包含了事件相关的时间、地点和主要参与者。因此，找出这些句子可以提高抽取论元的准确率。事件类型检测模块，模型分析了文档中各句子的重要性，据此，可以选出top-k个相关性较高的句子，候选句子中提及的实体作为候选实体。

有向无环图的路径扩展子任务为多个二分类问题的集合，即预测所有候选实体的扩展或不扩展。为了能够感知当前路径状态、历史上下文和当前论元角色，将上下文记忆向量$m$和实体向量$e^b$拼接起来，使用Transformer对其进行编码，促进上下文感知推理。从Transformer的输出中抽取相应的实体向量表示$e^r$，并在$e^r$上堆叠一个线性分类器来进行路径扩展分类。

$$e^r = \text{Transformer}(e^b, [m; e^b]) \tag{14}$$

$$p_{vr}^a = \text{softmax}(\text{FC}(e^r)) \tag{15}$$

其中，$p_{vr}^a$表示对于论元角色v，各候选实体r的预测概率。

每扩展一个节点，记忆向量更新一次，更新方式如下：

$$m' = [m; e^b] \tag{16}$$

论元抽取的损失函数为：

$$L_{ext} = -\sum_{v=1}^{V} \sum_{r=1}^{Q} \log p_{vr}^a \tag{17}$$

其中，V表示事件类型对应的论元角色个数，Q表示候选实体的个数。


(c) 2024 中国中文信息学会计算语言学专业委员会

### 3.5 基于多任务学习的联合优化

利用多任务学习的方式，联合3.2节获取实体提及、3.3节判断事件类型和3.4节抽取事件论元三者的损失，得到最终损失$L$：

$$L = \lambda_1 L_{ner} + \lambda_2 L_{dect} + \lambda_3 L_{ext} \tag{18}$$

其中，$\lambda_1, \lambda_2, \lambda_3$是超参数。

## 4 实验和分析

### 4.1 数据集

本文在两个公开的中文数据集ChFinAnn(Zheng et al., 2019)和DuEE-Fin[0]上进行了实验，ChFinAnn包含5个事件类型，24种不同的论元角色。DuEE-Fin包含13个事件类型，61种不同的论元角色。两个数据集的详情及划分如表1所示。

| | ChFinAnn | DuEE-Fin |
|---|---|---|
| 训练集 | 25632 | 6011 |
| 验证集 | 3204 | 1004 |
| 测试集 | 3204 | 1171 |
| 单事件 | 71% | 67% |
| 多事件 | 29% | 33% |

表 1: ChFinAnn和DuEE-Fin数据集详情

### 4.2 评价指标和参数设置

文档级事件抽取的最终目标是为每个事件的论元角色分配正确的实体，因此，在论元级别计算预测事件与真实事件标签的微平均的精确率P、召回率R和F1。本文设置最大句子数量和最大句子长度为64和128。为保证实验对比的公平性，本文选择了与对比模型MMR等模型相同的基座BERT-Base Chinese[1](Devlin et al., 2019)，Transformer的隐藏层和前馈层的维度分别是768和1024。训练过程中，设置$\lambda_1 = 0.05, \lambda_2 = \lambda_3 = 0.95$，学习率为$5e^{-5}$，梯度累积步数为8，使用Adam优化器，并在NVIDIA A100 GPU(40G)上训练，选择验证集上分数最高的轮次用于测试集的评估。

### 4.3 对比基线模型

本文选择了以下代表性的模型和大语言模型进行对比：

• DCFEE(Yang et al., 2018)：该模型从文档的中心句抽取论元，针对单事件和多事件抽取任务提出了两个变体DCFEE-O和DCFEE-M。

• Doc2EDAG(Zheng et al., 2019)：该模型利用基于实体的有向无环图实现文档级多事件抽取，Greedy-Doc是一个利用贪婪策略填充事件表的基础模型。

• GIT(Xu et al., 2021)：该模型利用异质图建模句子和实体提及之间的关系，并提出了一种跟踪机制捕获事件间的依赖。

• DE-PPN(Yang et al., 2021)：该模型利用Transformer编码句子和实体，用多粒度解码器并行地抽取所有事件。

• PTPCG(Zhu et al., 2022)：该模型是一个轻量级模型，构造了伪触发词的修剪完全图用于非自回归解码。

• ReDEE⋆(Liang et al., 2022)：为保证本文与对比实验的公平性，将ReDEE模型的参数gradient_accumulation_steps和train_batch_size分别设置为4和4，记作ReDEE⋆模型。

• MMR(Yang et al., 2022)：该模型将文档级事件抽取任务视为机器阅读理解任务，把论元抽取转化为从文档中识别论元，并基于MRC捕获角色与文档上下文的依赖关系。

• TFECI(Chen et al., 2023)：该模型提出了一种选择伪触发词的策略，利用实体共指和实体的缩写信息提高抽取性能。

[0] https://aistudio.baidu.com/datasetdetail/186939

[1] https://github.com/google-research/bert


(c) 2024 中国中文信息学会计算语言学专业委员会

• Zou(Zou et al., 2023)：该研究提出了一种捕捉实体信息与文档信息的编码方法，并在解码过程中预测出每种事件类型的事件个数。

• ChatGLM-6B[2]:清华大学发布的具有60亿参数的生成式大语言模型，采用P-Tuning的微调方式。

• Qwen-1.8B[3]:阿里云研发的具有18亿参数规模的大语言模型，本文对其进行全参数微调。

## 4.4 实验

### 4.4.1 实验结果和分析

为了评估所提模型的整体性能，将本文方法与第4.3节介绍的对比基线模型在ChFinAnn数据集上进行对比实验，实验结果见表2。

| Model | F1 for each event type | | | | | overall | | |
|---|---|---|---|---|---|---|---|---|
| | EF | ER | EU | EO | EP | P | R | F1 |
| DCFEE-O | 46.7 | 80.0 | 47.5 | 46.7 | 56.1 | 67.7 | 54.4 | 60.3 |
| DCFEE-M | 42.7 | 73.3 | 45.8 | 44.6 | 53.8 | 58.1 | 55.2 | 56.6 |
| Greedy-Doc | 57.7 | 79.4 | 51.2 | 50.0 | 54.2 | 80.4 | 49.1 | 61.0 |
| Doc2EDAG | 71.0 | 88.4 | 69.8 | 73.5 | 74.8 | 80.3 | 75.0 | 77.5 |
| GIT | 73.4 | 90.8 | 74.3 | 76.3 | 77.7 | 82.3 | **78.4** | 80.3 |
| DE-PPN | 73.5 | 87.4 | **74.4** | 75.8 | **78.4** | 83.7 | 76.4 | 79.9 |
| PTPCG | — | — | — | — | — | **88.2** | 69.1 | 79.4 |
| ReDEE⋆ | 63.4 | 88.6 | 61.4 | 69.6 | 74.0 | 84.0 | 69.6 | 76.1 |
| MMR | 71.3 | 88.7 | 72.4 | **76.9** | 78.0 | 83.8 | 76.3 | 79.9 |
| TFECI | **74.1** | 91.6 | 70.9 | 71.4 | 76.7 | 83.3 | 76.0 | 79.5 |
| Zou | 72.4 | 90.4 | 72.6 | 73.3 | 74.8 | 81.6 | 75.1 | 78.2 |
| ChatGLM-6B | — | — | — | — | — | 34.2 | 30.8 | 32.4 |
| Qwen-1.8B | — | — | — | — | — | 46.8 | 46.7 | 46.7 |
| IESLHI(ours) | 70.8 | **92.9** | 72.5 | 74.0 | 78.1 | 86.3 | 75.9 | **80.7** |

⋆：本文用(Liang et al., 2022)公开的代码复现的实验结果(下同)。P、R、F1分别表示Presion(%)、Recall(%)、F1-score(%)(下同)。股权冻结(EF)、股权回购(ER)、股权减持(EU)、股权增持(EO)、股权质押(EP)。

表 2: 模型在ChFinAnn数据集上的实验结果

由表2可知：

（1）本模型在F1指标上的整体表现均优于对比基线，相比最优的对比模型，本模型取得了0.4%的提升。特别地，相比Doc2EDAG模型，本模型进一步度量了句子与事件类型的关联差异性，其能够约束候选实体的抽取范围，进而缓解论元分散，并取得了3.2%的提升。

（2）在事件类型ER和EP中，本模型取得了最优或次优的表现，表明通过度量句子与事件类型的关联差异性，可以抽取出文档中与事件类型关联性较强的句子，从而有效地约束候选实体的识别范围。

（3）对于事件类型EF、EU和EO，本文方法不及GIT、DE-PPN和MMR，可能原因是存在事件类型如股权增持(EO)等事件类型的大部分文档包含40多个句子，论元在文档中较为分散、且跨度较大。而本文未建立句子之间的关系，导致模型在这些事件类型上的抽取性能不理想。

为了进一步验证模型的性能，将本文方法与第4.3节介绍的对比基线模型在DuEE-Fin数据集上进行对比实验，实验结果见表3。

由表3可知：

（1）本模型相比最优的基线模型在F1指标上提高了0.3%，其在大多数事件类型上取得了最优或次优的结果，表明通过捕获事件类型与关联性较强的句中实体的关系有助于提高事件抽取性能，验证了所提模型的有效性。

（2）对于一些事件类型，本文方法次于GIT和ReDEE，可能原因是本模型没有对句间关系建模，而解除质押(RP)、亏损(Lo)等事件类型的文档长度较长，无法较好地捕获句间依赖关系。

[2]https://github.com/THUDM/ChatGLM-6B

[3]https://github.com/QwenLM/Qwen


(c) 2024 中国中文信息学会计算语言学专业委员会

| Model | F1 for each event type | | | | | | | | | | | | | overall | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | WB | Lo | CA | EB | CF | GP | EU | EO | ER | Vi | RP | Pl | EC | P | R | F1 |
| DCFEE-O | 54.0 | 65.4 | 44.0 | 27.3 | 58.2 | 42.0 | 48.8 | 53.9 | 76.7 | 32.9 | 63.3 | 58.3 | 40.6 | 59.8 | 55.5 | 57.6 |
| DCFEE-M | 49.2 | 68.0 | 40.4 | 28.4 | 51.2 | 35.1 | 42.3 | 45.9 | 74.0 | 51.0 | 55.8 | 56.4 | 37.4 | 50.2 | 55.5 | 52.7 |
| Greedy-Doc | 53.7 | 71.8 | 49.5 | 41.1 | 61.3 | 42.1 | 49.7 | 57.4 | 74.4 | 29.2 | 60.8 | 50.5 | 39.4 | 66.0 | 50.6 | 57.3 |
| Doc2EDAG | 60.0 | **78.3** | 50.6 | 40.1 | 63.2 | 51.5 | 50.7 | 52.9 | 83.7 | 51.2 | 64.8 | 61.7 | 51.2 | 73.7 | 59.8 | 66.0 |
| GIT | 58.8 | 77.6 | 56.6 | 44.7 | 68.5 | **55.1** | 58.8 | **71.2** | 86.4 | 45.0 | 66.4 | 71.3 | **53.8** | 75.4 | 61.4 | 67.7 |
| DE-PPN | 50.7 | 62.7 | 41.3 | 21.4 | 36.3 | 23.0 | 32.9 | 31.3 | 67.8 | 25.8 | 42.1 | 36.3 | 23.4 | 69.0 | 33.5 | 45.1 |
| PTPCG | — | — | — | — | — | — | — | — | — | — | — | — | — | 71.0 | 61.7 | 66.0 |
| ReDEE ⋆ | 65.9 | 75.9 | 56.1 | **48.7** | 69.4 | 51.0 | 64.1 | 56.4 | **87.5** | 51.6 | **71.5** | 70.3 | 52.5 | **79.8** | 60.9 | 69.1 |
| ChatGLM-6B | — | — | — | — | — | — | — | — | — | — | — | — | — | 38.7 | 36.6 | 37.6 |
| Qwen-1.8B | — | — | — | — | — | — | — | — | — | — | — | — | — | 41.5 | 38.9 | 40.1 |
| IESLHI(ours) | **68.4** | 76.7 | **56.8** | 47.7 | **71.9** | 53.1 | **65.2** | 65.0 | 81.6 | **56.6** | 71.0 | **75.8** | 52.8 | 71.0 | **67.8** | **69.4** |

公司上市(GP)、股东减持(EU)、股东增持(EO)、企业收购(CA)、企业融资(CF)、股份回购(ER)、高管变动(EC)、解除质押(RP)、企业破产(EB)、亏损(Lo)、被约谈(Vi)、中标(WB)、质押(Pl)。

表 3: 模型在DuEE-Fin数据集上的实验结果

（3）各模型在DuEE-Fin数据集上的整体表现较差，主要原因是ChFinAnn的训练集数据量是DuEE-Fin的5倍，而DuEE-Fin的事件类型是ChFinAnn的2.6倍。

另外，综合表2和表3可以观察到：本文方法与大语言模型的性能对比，ChatGLM和Qwen的微调结果在特定领域的文档级事件抽取任务上表现不佳。在ChatGLM和Qwen微调过程中，构造了提示Prompt4=“[文档]，请判断文档包含的事件类型并抽取相应的论元：[[事件类型1，论元角色1：论元1，论元角色2：论元2],[事件类型2，论元角色3：论元3，论元角色4：论元4]]”，尽管已经在训练数据上对所对比的大模型进行了训练和微调，但是它们处理语义理解和推理等复杂任务的能力依然有限，进一步表明为复杂任务设计相应的模型是有必要的。本文方法既利用了大语言模型丰富的背景语义信息，建立了语义扩展的实体抽取与表示模块，并通过获取文本的扩展语义，对事件进行了有效的表示。在此基础上，又建立了基于关联差异性的事件类型检测模块，用于捕获长文本中与事件类型关联性较高的句子，以约束候选论元的抽取范围。而大模型无法捕获长文本中与事件类型高度相关的句子，进而无法约束候选论元的抽取范围，容易导致多个事件-论元间对应关系出现混乱的问题。

### 4.4.2 单事件和多事件抽取分析

为了评估所提模型在单事件和多事件上的表现，本文将ChFinAnn数据集中涉及单个事件和多个事件的数据分别进行分析，结果如表4所示。

| Model | EF | | ER | | EU | | EO | | EP | |
|---|---|---|---|---|---|---|---|---|---|---|
| | S. | M. | S. | M. | S. | M. | S. | M. | S. | M. |
| DCFEE-O | 55.7 | 38.1 | 83.0 | 55.5 | 52.3 | 41.4 | 49.2 | 43.6 | 62.4 | 52.2 |
| DCFEE-M | 45.3 | 40.5 | 76.1 | 50.6 | 48.3 | 43.1 | 45.7 | 43.3 | 58.1 | 51.2 |
| Greedy-Doc | 74.0 | 40.7 | 82.2 | 50.0 | 61.5 | 35.6 | 63.4 | 29.4 | 78.6 | 36.5 |
| Doc2EDAG | 79.7 | 63.3 | 90.4 | 70.7 | 74.7 | 63.3 | 76.1 | 70.2 | 84.3 | 69.3 |
| GIT | 81.9 | **65.9** | 93.0 | 71.7 | **82.0** | 64.1 | **80.9** | 70.6 | 85.0 | **73.5** |
| DE-PPN | **82.1** | 63.5 | 89.1 | 70.5 | 79.7 | **66.7** | 80.6 | 69.6 | 88.0 | 73.2 |
| ReDEE ⋆ | 77.9 | 49.1 | 90.8 | 67.9 | 71.6 | 46.7 | 74.8 | 62.8 | 83.6 | 68.2 |
| MMR | 81.2 | 61.8 | 89.8 | 70.1 | 77.9 | 65.4 | 80.8 | **71.7** | 86.2 | 72.6 |
| Zou | 81.2 | 64.8 | 92.6 | 73.6 | 78.4 | 65.4 | 80.3 | 65.5 | **89.6** | 66.2 |
| IESLHI(ours) | 81.5 | 60.6 | **93.6** | **75.8** | 80.3 | 62.2 | 79.5 | 67.3 | 86.4 | 73.3 |

S.表示单事件，M.表示多事件。

表 4: 模型在ChFinAnn数据集上的单事件和多事件比较

由表4可知:

（1）在大多数事件类型的单事件中，本模型取得了最优或次优的表现。各模型单事件的F1整体都高于多事件的F1，体现出多事件抽取任务是有一定难度的。

（2）事件类型ER、EP的多事件中，本模型表现较好，表明所提模型通过判断不同事件类型的相关句子，并分别对每个事件类型进行路径扩展以抽取论元的策略可以提高多事件抽取的性能。


(c) 2024 中国中文信息学会计算语言学专业委员会

（3）本模型在事件类型EF、EU、EO的多事件中表现并不理想，可能原因是多个事件间存在事件类型重叠的情况，对解码路径造成一定的干扰，导致出现漏判或错判的情况。

### 4.4.3 消融实验

为了证明所提模型每个模块的有效性，设计了三种变体模型进行消融实验。1）移除事件类型标签和论元角色标签的解释(-标签解释)；2）移除判断事件类型与句子的关联差异性(-关联差异性)；3）移除文档的语义扩展(-语义扩展)。

| | ChFinAnn | | | DuEE-Fin | | |
|---|---|---|---|---|---|---|
| | P | R | F1 | P | R | F1 |
| IESLHI | **86.3** | **75.9** | **80.7** | **71.0** | **67.8** | **69.4** |
| -标签解释 | 88.7 | 70.5 | 78.6 | 67.3 | 68.3 | 67.8 |
| -关联差异性 | 86.5 | 72.9 | 79.1 | 65.7 | 70.8 | 68.2 |
| -语义扩展 | 87.5 | 72.7 | 79.4 | 69.4 | 68.1 | 68.7 |

表 5: IESLHI在ChFinAnn和DuEE-Fin数据集的消融实验

表5显示了移除任一模块，模型的性能都有所下降，说明每个模块对所提模型都有积极作用。可以观察到：

（1）移除标签解释在ChFinAnn和DuEE-Fin上F1分别下降了2.1%和1.6%，表明利用大模型生成的标签解释可以在解码图中发挥较大作用，帮助模型判断每一论元角色时正确识别相应的实体。

（2）移除关联差异性模块在ChEinAnn和DuEE-Fin上F1分别下降了1.6%和1.2%，表明度量与事件类型相关的句子的权重对模型性能提升也有较大的贡献，且该组件可以有效约束候选实体的产生范围，避免过多实体导致的信息冗余。

（3）移除语义扩展模块在ChEinAnn和DuEE-Fin上F1分别下降了1.3%和0.7%，显示出文档的语义扩展也可以对模型做出比较稳定的贡献，该语义扩展组件可以丰富句子的语义，为命名体识别任务提供帮助。

为了进一步分析各个模块对模型性能的影响，本文选取了一个多事件论元抽取的例子，如图3所示。根据图中IESLHI列的抽取结果，所提模型可以正确地抽取事件类型为“股权质押”的两个事件包含的所有论元，通过大语言模型对文本的解释和扩展，可以丰富文本信息，有助于正确地识别论元。




图 3: 多事件文档示例，橙色标注表示实体，绿色框线标注表示论元

对于移除标签解释或语义扩展模块后，存在对文档的理解偏差，导致论元抽取的缺失或误判，如将事件1中的开始时间“2018年9月25日”预测为“2018年9月20日”。而移除关联差异性模块

后，实体较多引入的冗余信息，导致论元抽取错误，如将事件2中的总持有股份“16679436股”预测为“39197988股”。

#### 4.4.4 关联差异性影响分析

为了评估关联差异性对模型性能的影响，本文选择与事件类型相关的句子个数K分别为3、6、8、10时在ChFinAnn数据集进行实验，实验结果见表6。

| | P | R | F1 |
|---|---|---|---|
| K = 3 | 89.2 | 51.2 | 65.1 |
| K = 6 | 88.1 | 73.7 | 80.2 |
| K = 8 | 86.3 | 75.9 | 80.7 |
| K = 10 | 85.8 | 74.7 | 79.9 |

表 6: 不同K值时IESLHI在ChFinAnn数据集上的比较

根据研究(Xu et al., 2021)，该数据集的每个文档平均由20个句子，912个token组成，一个事件记录大约涉及6个句子。由表6可以观察到：选择与事件类型关联性较强的句子个数K为8时，几乎可以包含文档中完整的事件记录所在的句子，避免了候选实体的缺失以及过多命名实体带来的信息冗余，因此，模型整体表现优异，P、R、F1值的表现较为稳定。

## 5 结论

本文针对文档级事件抽取任务中的论元分散和多事件提及问题，提出了融合扩展语义与标签层次信息的文档级事件抽取模型。该模型基于大语言模型对文档信息和标签信息进行语义扩展，利用关联差异性的判断选出了文档中与事件类型相关性较高的句子，针对每个事件类型，基于有向无环图从候选实体中抽取论元。在ChFinAnn和DuEE-Fin数据集上的实验表明，相比基线对比模型，所提模型在F1指标上分别提升了0.4%和0.3%，验证了通过关联差异性的判断可以选取与事件类型相关性较高的句子中的实体作为候选实体，有效地缓解论元分散；对每个事件类型分别进行路径扩展可以提升多事件提及的性能。下一步，我们将考虑基于异质图融入实体与句子的关系以捕获句子之间的长距离依赖关系，并进一步整合这些信息。

## 参考文献

Fumin Chen, Xu Wang, Xiaohui Liu, and Dezhong Peng. 2023. A trigger-free method enhanced by coreference information for document-level event extraction. In *Proceeding of the 2023 International Joint Conference on Neural Networks*, pages 1–8.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Li Du, Xiao Ding, Yue Zhang, Ting Liu, and Bing Qin. 2022. A graph enhanced BERT model for event prediction. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2628–2638.

Guanhua Huang, Runxin Xu, Ying Zeng, Jiaze Chen, Zhouwang Yang, and Weinan E. 2023. An iteratively parallel generation method with the pre-filling strategy for document-level event extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10834–10852.

Sha Li, Heng Ji, and Jiawei Han. 2021. Document-level event argument extraction by conditional generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 894–908.

Qian Li, Jianxin Li, Jiawei Sheng, Shiyao Cui, Jia Wu, Yiming Hei, Hao Peng, Shu Guo, Lihong Wang, Amin Beheshti, and Philip S. Yu. 2022. A survey on deep learning event extraction: Approaches and applications. *IEEE Transactions on Neural Networks and Learning Systems*, 35(5):1–21.


(c) 2024 中国中文信息学会计算语言学专业委员会

Yuan Liang, Zhuoxuan Jiang, Di Yin, and Bo Ren. 2022. RAAT: Relation-augmented attention transformer for relation modeling in document-level event extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4985–4997.

Jian Liu, Chen Liang, Jinan Xu, Haoyan Liu, and Zhe Zhao. 2023a. Document-level event argument extraction with a chain reasoning paradigm. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 9570–9583.

Wanlong Liu, Shaohuan Cheng, Dingyi Zeng, and Qu Hong. 2023b. Enhancing document-level event argument extraction with contextual clues and role relevance. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12908–12922.

Yaojie Lu, Hongyu Lin, Jin Xu, Xianpei Han, Jialong Tang, Annan Li, Le Sun, Meng Liao, and Shaoyi Chen. 2021. Text2Event: Controllable sequence-to-structure generation for end-to-end event extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 2795–2806.

Yubo Ma, Zehao Wang, Yixin Cao, Mukai Li, Meiqi Chen, Kun Wang, and Jing Shao. 2022. Prompt for extraction? PAIE: Prompting argument interaction for event argument extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 6759–6774.

Amir Pouran Ben Veyseh, Minh Van Nguyen, Franck Dernoncourt, Bonan Min, and Thien Nguyen. 2022. Document-level event argument extraction via optimal transport. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1648–1658.

Yubing Ren, Yanan Cao, Ping Guo, Fang Fang, Wei Ma, and Zheng Lin. 2023. Retrieve-and-sample: Document-level event argument extraction via hybrid retrieval augmentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 293–306.

Jiawei Sheng, Shu Guo, Bowen Yu, Qian Li, Yiming Hei, Lihong Wang, Tingwen Liu, and Hongbo Xu. 2021. CasEE: A joint learning framework with cascade decoding for overlapping event extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 164–174.

Qizhi Wan, Changxuan Wan, Keli Xiao, Dexi Liu, Chenliang Li, Bolong Zheng, Xiping Liu, and Rong Hu. 2023. Joint document-level event extraction via token-token bidirectional event completed graph. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 10481–10492.

Xinyu Wang, Lin Gui, and Yulan He. 2023. Document-level multi-event extraction with event proxy nodes and hausdorff distance minimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 10118–10133.

Runxin Xu, Tianyu Liu, Lei Li, and Baobao Chang. 2021. Document-level event extraction via heterogeneous graph-based interaction model with a tracker. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 3533–3546.

Runxin Xu, Peiyi Wang, Tianyu Liu, Shuang Zeng, Baobao Chang, and Zhifang Sui. 2022. A two-stream AMR-enhanced model for document-level event argument extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5025–5036.

Hang Yang, Yubo Chen, Kang Liu, Yang Xiao, and Jun Zhao. 2018. DCFEE: A document-level Chinese financial event extraction system based on automatically labeled training data. In *Proceedings of ACL 2018, System Demonstrations*, pages 50–55.

Hang Yang, Dianbo Sui, Yubo Chen, Kang Liu, Jun Zhao, and Taifeng Wang. 2021. Document-level event extraction via parallel prediction networks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 6298–6308.

Hang Yang, Yubo Chen, Kang Liu, Jun Zhao, Zuyu Zhao, and Weijian Sun. 2022. Multi-turn and multi-granularity reader for document-level event extraction. *Association for Computing Machinery*, 22(2):1–16.


(c) 2024 中国中文信息学会计算语言学专业委员会

Shun Zheng, Wei Cao, Wei Xu, and Jiang Bian. 2019. Doc2EDAG: An end-to-end document-level framework for Chinese financial event extraction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 337–346.

Tong Zhu, Xiaoye Qu, Wenliang Chen, Zhefeng Wang, Baoxing Huai, Nicholas Yuan, and Min Zhang. 2022. Efficient document-level event extraction via pseudo-trigger-aware pruned complete graph. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4552–4558.

Jinan Zou, Yanxi Liu, Yuankai Qi, Haiyao Cao, Lingqiao Liu, and Javen Qinfeng Shi. 2023. A generative approach for comprehensive financial event extraction at the document level. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, page 323–330.

李治. 2022. 基于实时事件侦测的兴趣点推荐系统研究. 数据分析与知识发现, page 114.

CCL 2024

# 融合领域词汇扩充的低资源法律文书命名实体识别

**帕尔哈提·吐拉江[ab], 孙媛媛*[a], 蔡艾辰[a], 王艳华[c], 林鸿飞[a]**
[a]大连理工大学，计算机科学与技术学院，大连，116024
[b]新疆师范大学，计算机科学技术学院，乌鲁木齐，830054
[c]中国人民解放军空军通信士官学校，大连，710038
(prht,18941991166)@mail.dlut.edu.cn,(syuan,hflin)@dlut.edu.cn
wangyanhua818@foxmail.com

## 摘要

目前基于预训练语言模型的司法领域低资源法律文书命名实体识别研究主要面临两个问题：(1)在低资源语言中，如维吾尔语，法律文书相关的语料极其有限，这种语料资源稀缺限制了基于预训练语言模型的训练和性能。(2)法律文书中使用的专业术语不仅复杂且特定，新的法律术语和概念的出现使得现有的模型难以适应。针对上述问题，本文基于多语言预训练模型mBERT，通过领域词汇扩充及模型微调的方法，提升了模型在维吾尔语法律文书命名实体识别任务的性能。本文首先整理并构建维吾尔语司法领域专业词汇列表，并将其添加到mBERT模型的词汇表中。随后，在人工标注的维吾尔语法律文书命名实体数据集UgLaw-NERD上进行模型微调，验证了该方法的有效性。实验结果表明，相比于仅使用mBERT进行微调的基线模型，融合领域词汇扩充的模型在命名实体识别任务上F1得分提升至89.72%，较基线提高了7.39%。此外，本文还探讨了不同领域词汇扩充量对模型命名实体识别性能的影响，结果显示，领域词汇扩充增强了预训练模型在处理维吾尔语任务中的表现。这些结论为其他低资源语言在司法领域开展基于预训练模型的自然语言处理研究提供了有益的参考。

**关键词：** 司法领域； 词汇扩充； 预训练语言模型； 低资源； 维吾尔语； 命名实体识别

# Named Entity Recognition for Low-Resource Legal Documents Using Integrated Domain Vocabulary Expansion

**PAERHATI Tulajiang[ab],Yuanyuan Sun* [a], Aichen Cai[a], Yanhua Wang[c], Hongfei Lin[a]**
[a]School of Computer Science and Technology,Dalian University of Technology,Dalian,116024
[b]College of Computer Science and Technology, Xinjiang Normal University, Urumqi,830054
(prht,18941991166)@mail.dlut.edu.cn,(syuan,hflin)@dlut.edu.cn
wangyanhua818@foxmail.com

## Abstract

Current research in named entity recognition for low-resource legal documents, based on pre-trained language models, primarily faces two challenges: (1) In low-resource languages, such as Uyghur, relevant legal document corpora are extremely scarce, limiting the training and performance of models based on pre-trained language models. (2) The use of complex and specific legal terminology in legal documents, along with the emergence of new legal terms and concepts, makes it difficult for existing models to adapt. In response to these challenges, this paper leverages the multilingual pre-trained model mBERT and enhances its performance on the Uyghur legal document named entity recognition task through domain vocabulary expansion and model fine-tuning.

* 通讯作者

第二十三届中国计算语言学大会论文集，第431页-第441页，太原，中国，2024年7月25日至28日。
卷1：主会论文
(c) 2024 中国中文信息学会计算语言学专业委员会

This study begins by compiling and constructing a list of Uyghur legal domain-specific vocabulary and integrating it into the mBERT model's vocabulary. Following this, the model is fine-tuned on the manually annotated Uyghur legal document named entity dataset, UgLaw-NERD, validating the effectiveness of this approach. Experimental results indicate that the domain vocabulary-expanded model achieves an F1 score of 89.72%, an increase of 7.39% over the baseline model that only utilizes fine-tuning with mBERT. Additionally, this paper explores the impact of varying amounts of domain vocabulary expansion on model performance in named entity recognition, revealing that domain vocabulary expansion enhances the pre-trained model's capability in processing Uyghur language tasks. These findings offer valuable insights for other low-resource languages embarking on natural language processing research in the legal field based on pre-trained models.

**Keywords:** Legal Domain , Vocabulary Expansion , Pre-trained Language Models , Low-resource , Uyghur , Named Entity Recognition

## 1 引言

在司法领域，命名实体指法律文书中具有特定司法属性的名词和短语(李春楠et al., 2021; Chen et al., 2020)，主要包括涉案时间、地点、相关人员(如犯罪嫌疑人、被害人等)、组织机构、罪名以及涉案金额等。针对司法命名实体的识别研究不仅是法律文书案件情节关系抽取、构建司法知识图谱的基础工作，而且对法律问答、刑期预测、类案推荐等司法应用的实际落地具有至关重要的作用。

基于统计的机器学习方法在司法领域命名实体识别研究中，Leitner等人(Leitner et al., 2019)手工设计了多组特征，并将这些特征与CRF模型结合使用，以完成法律文书的命名实体识别。虽然CRF模型能根据词汇之间的依赖关系及其与标签的关系来建模概率分布，但在处理法律文书时受到上下文考虑受限的限制，通常只能处理局部上下文而难以捕捉长距离依赖关系。并且模型性能高度依赖于手工设计的特征工程，难以做到全面和准确。随着深度学习的发展，很多研究(Xu et al., 2021; Bonifacio et al., 2020)已经将BERT等预训练语言模型（PLM）应用到了司法领域，极大地提高了法律文本的特征编码质量，以及对法律文本的理解和信息抽取能力。然而，目前基于预训练模型针对司法领域命名实体识别任务的探索主要集中在资源丰富的语言上，对维吾尔语等低资源语言的研究还处于初级阶段，公开的多语言预训练语言模型不能很好地应用于低资源语言命名实体识别任务上。

命名实体识别技术在不同语言环境中的效果各不相同，尤其是在资源匮乏的低资源语言和语言结构复杂的非主流语言中面临显著挑战。以维吾尔语为例，早期的维吾尔语命名实体识别始于基于词典和规则(Collins and Singer, 1999)、基于统计机器学习(Isozaki and Kazawa, 2002)的方法的探索，主要应用基于规则和条件随机场(CRF)(Lafferty et al., 2001)方法对人名、地名、组织机构等实体类型进行识别。随着深度学习技术的快速进步，研究和应用逐渐转向基于深度学习的端到端方法。吾守尔等人(吾守尔et al., 2018)基于BiLSTM-CNN-CRF的神经网络模型，在资源稀缺环境下提高了维吾尔语命名实体识别的性能。AzmatAnwar等人(Anwar et al., 2020)应对在资源匮乏的维吾尔语上进行神经网络命名实体识别的挑战，提出了基于神经机器翻译的命名实体标注转移方法。虽然基于深度神经网络的方法(王路路et al., 2019)在维吾尔语命名实体识别任务上达到了90%以上的F1分数，但这些任务主要针对通用领域的命名实体类型（如人名、地名和组织机构）的识别。相对于通用领域，司法领域命名实体识别任务所关注的实体类型粒度小、分类多，包含法律专业术语，再加上维吾尔语等低资源语言司法领域语料的稀缺性和语法规则的复杂性，给低资源语言法律文书的命名实体识别研究带来了额外的挑战。

对低资源语言，哈工大讯飞联合实验室Yang等人(Yang et al., 2022)以多语言预训练模型XLM-R为基座，训练了第一个面向中国少数民族语言的预训练语言模型CINO，并在文本分类任务上进行微调，提高了模型文本分类任务上的能力。Deng等人(Deng et al., 2023)提出的预

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

训练模型MiLMo在CINO-base-v2模型的基础上，进一步提升了模型性能。虽然多语言预训语言练模型在低资源语言分类任务上展现出了优异的性能，但以上研究都是面向通用领域的分类任务，并没有对特定领域的命名实体识别任务开展相关研究。

从上述研究工作来看，基于预训练语言模型的面向司法领域低资源语言法律文书命名实体识别研究主要面临两个问题：(1)在低资源语言中，如维吾尔语，法律文书相关的语料极其有限，这种语料资源稀缺限制了基于预训练语言模型的训练和性能。(2)法律文书中使用的术语不仅复杂且特定。法律法条的新增，修订、修正和修改，使得之前针对法律文本训练的预训练模型因缺少新增法律术语和概念的专业领域词汇，可能会造成下游任务的性能下降。

鉴于上述问题，本文提出了一种融合领域词汇扩充和预训练模型微调的方法，实现低资源语言法律文书的命名实体识别。本文采用了多语言预训练模型mBERT(Pires et al., 2019)作为基线，融合词汇扩充和模型微调的策略，提高维吾尔语法律文书中命名实体识别的性能。具体地，首先从中国裁判文书网[1]收集维吾尔语法律文书，经过数据处理，构建专门针对司法领域的维吾尔语词汇表。然后对预训练模型的词汇表进行扩充，并对模型的嵌入层进行调整。最后，通过数据增强和人工标注构建的维吾尔语法律文书命名实体数据集UgLaw-NERD对模型进行微调，实现对维吾尔语法律文书命名实体的识别。本文主要贡献总结如下:

（1）本文构建了专门针对司法领域的维吾尔语法律文书命名实体识别数据集UgLaw-NERD[2]，该数据集包含9个细粒度司法实体类型、3269句子和96745个token。

（2）提出了一种融合预训练模型的词汇扩充策略。通过结合词汇扩充与模型微调，该方法提高了预训练模型在维吾尔语法律文书命名实体识别的表现。

（3）本文在UgLaw-NERD数据集上进行实验，与现有方法相比，本文提出的方法展现了性能的提升，也证实了UgLaw-NERD数据集在维吾尔语法律文书命名实体识别领域的应用价值。

本文其余部分组织如下：第2章主要介绍低资源语言命名实体识别相关研究。第3章介绍数据集的开发过程和注释指南。第4章介绍本文提出的方法。第5章主要介绍实验设置、实验结果及参数分析。第6章为总结与未来研究方向的探讨。

## 2 相关工作

现代维吾尔语中有32个字母，每个单词都由字母拼写而成。维吾尔语单词的词形结构复杂，词尾通过接不同的词缀来实现其语法功能。针对维吾尔语的命名实体识别的研究中，木合塔尔等人(木合塔尔·艾尔肯et al., 2013)通过手动建立基于新疆维吾尔自治区的地名词典库等，提出了基于规则的维吾尔语地名识别技术。艾斯卡尔等人(艾斯卡尔·肉孜et al., 2013)实现了基于条件随机场的维吾尔人名识别方法。加日拉等人(加日拉·买买提热衣木et al., 2014)采用统计和规则相结合的混合策略，提出了一种维吾尔人名的自动识别方法。麦合甫热提等人(麦合甫热提et al., 2014)则从维吾尔语的语言特点出发，对维吾尔语中机构名称进行分类和形式化表示，并设计了基于状态转移原理的维吾尔文机构名识别算法。买合木提·买买提等人(买合木提·买买提et al., 2019)通过分析维吾尔文地名，提出了一种结合条件随机场和规则的识别方法。

随着深度学习技术的快速进步，研究和应用已逐渐转向基于深度学习的端到端方法。利用卷积神经网络(CNN)(Chen, 2015)和循环神经网络(RNN)(Zaremba et al., 2014)等结构，从文本序列中抽取隐含特征，再通过应用条件随机场(CRF)来确定最优实体序列。该方法直接对原始数据进行深度学习建模和特征提取，不仅避免了传统方法中依赖复杂的人工特征工程，还有效解决了数据稀疏性导致的维度灾难问题。王路路等人(王路路et al., 2018)针对维吾尔文命名实体识别无法利用未标注预料的问题，提出了基于CRF和半监督学习的方法，并进一步结合注意力机制与BiLSTM-CRF(王路路et al., 2019)，提升了维吾尔语命名实体识别识别率。

在构建维吾尔语命名实体标注语料方面，艾斯卡尔等人(艾斯卡尔·肉孜et al., 2013)根据维吾尔人名特点收集了5258个句子，形成了维吾尔人名语料库。汪昆等人(汪昆et al., 2017)在人名、地名、机构名的一体化识别任务中构建了共有11257个标注句子的维吾尔语命名实体识别语料。Kahaerjiang等人(Abiderexiti et al., 2017)构建了第一个命名实体关系识别UyNeRel数据集。孙祥鹏等人(孔祥鹏et al., 2020)研究基于迁移学习的维吾尔语命名实体识别方法中构建了基于新闻预料标注的维吾尔语命名实体识别数据集。此外，除了通用领域的人名、地名、机构

[1]https://wenshu.court.gov.cn/
[2]https://github.com/PaErHaTi-DUTiR/UgLaw-NERD


(c) 2024 中国中文信息学会计算语言学专业委员会

名等实体类型之外，阿迪来等人(阿迪来et al., 2017)在对维吾尔语音乐实体识别研究中，构建了2400句含有音乐实体的数据集。维吾尔语作为低资源语言，只有少量公开的语料库可获取。

词汇扩充是指在现有预训练语言模型的基础上，增加特定领域或语言的词汇，以解决模型原有词汇表不能充分覆盖特定用途语言元素的问题。这一过程涉及将新词汇和表达方式整合到模型的词汇表中，从而扩展模型的语言理解和生成能力。通过词汇扩充，模型能够更好地处理领域特定的术语和表达，增强其在特定任务和环境中的应用性能。Chung等人(Chung et al., 2020)探讨了如何通过扩大词汇表来优化预训练的多语言模型，避免了使用大词汇表重新训练深度模型的需要。Tai等人(Tai et al., 2020)引入了一种通过在有限的训练资源下添加新的附加词汇来扩展BERT预训练模型的方法。Tanaka等人(Tanaka and Shinnou, 2022)提出为预训练的BERT模型添加词嵌入来扩展复合词词汇表的方法。

本文主要为解决在低资源环境中维吾尔语特定司法领域专业术语难以被预训练语言模型有效识别的问题，引入了领域词汇扩充的策略，采用多语言预训练语言模型在司法领域数据集上进行微调，在深层次上捕捉维吾尔语法律文本的特定语义和结构特点，优化模型的命名实体识别能力。本文融合领域词汇扩充和预训练模型微调的方法在低资源语言法律文书的命名实体识别研究，能够在一定程度上填补对这一领域的研究空白和资源建设的需求。同时，为其他低资源语言在司法领域的自然语言处理研究提供有益借鉴和参考。

## 3 数据集构建

### 3.1 数据来源

本文数据来源于中国裁判文书网，这是一个公开提供各类法律文书的权威平台。截止目前，中国裁判文书网已收入文书总量达6414万份，访问总量高达228亿次，访客来自全球210多个国家和地区，成为全球最大的裁判文书公开平台。中国裁判文书网公开提供了藏语、蒙古语、维吾尔语、哈萨克语和朝鲜语5种我国少数民族语言的法律文书，这为在司法领域开展多语言自然语言处理研究提供了数据资源。基于研究的需求和目标，本文收集整理了3921篇维吾尔语的法律文书，涵盖了民事、刑事、赔偿、执行以及行政等多种案件类型，覆盖了维吾尔语司法领域的广泛案例。本文以盗窃类案件作为研究对象，通过数据增强和人工标注，建立一个专用的维吾尔语法律文书命名实体数据集UgLaw-NERD，以支持对司法领域维吾尔语命名实体识别技术的研究，也为深入探索维吾尔语在司法领域的文本处理提供基础。

### 3.2 数据预处理

中国裁判文书网提供的民族语言裁判文书均是PDF文件，需要将PDF文档转化为可处理的文本。本文应用科大讯飞多语种印刷文字识别API[3]，实现PDF转换和OCR识别，完成了文本数据的清洗。另外还采用翻译增强，将中文法律文书翻译成维吾尔语扩展了数据集并增加多样性。通过规则识别和提取文书中的案件描述，最终，将内容划分为3269个独立的盗窃类事件提及，每个描述都是一个完整的信息单元，详细反映了案件的具体情况。通过以上的处理过程，从大量维吾尔语法律文书中提取出关键信息段落，为后续的命名实体识别和数据分析准备了数据资源。

### 3.3 数据标注

本文除了通用领域命名实体识别任务中常用的实体类型，如人名、时间、地点之外，增加了司法领域命名实体识别任务中对案件理解十分重要的特有实体类型。本文定义了9种司法领域命名实体识别类型，以更好的抽取维吾尔法律文书种的案件要素。实体类型及定义如表1所示。

数据标注过程中，本文采用BIO(Begin-Inside-Outside)(Thompson and Lascarides, 1999)标注方案来标注命名实体。该方案分为三种标签：B 代表实体的开始，I 代表实体的内部，而O则表示非实体部分。相比IOBES(Uchimoto et al., 2000)方案，BIO标注方案在本文实验过程中，提供了一个较好的平衡点，既能有效识别大部分实体类型，更适用于维吾尔语法律文书的实体识别任务，也便于模型学习和实现。图1展示了标注数据案例。

### 3.4 人工评估

四名母语为维吾尔语的法学专业学生经过接受为期两周的命名实体数据标注培训后，被分

[3]https://www.xfyun.cn/services/xf-printed-word-recognition


(c) 2024 中国中文信息学会计算语言学专业委员会

| 标签 | 定义 |
|---|---|
| **NHCS** | 案件中出现的嫌疑人，被告人的姓名等表示身份的名称。 |
| **NT** | 案件中所涉及的时间段和某一时刻的时间表达式。 |
| **NS** | 案件所涉及的地理位置信息，包括行政区地名，街巷名。 |
| **NCGV** | 案件中所涉及的被盗物品价值，形式包括贵金属货币，纸币，电子货币等。 |
| **NHVI** | 案件中的受害人，被害人的姓名等表示身份的名称。 |
| **NCSM** | 案件中所涉及的被盗的现金数额价值，包括¥，人民币，元等字符。 |
| **NASI** | 案件中被害人的财产损失，被盗物品等。 |
| **NO** | 案件涉及的行政机构、企业机构及民间团体组织。 |
| **NATS** | 嫌疑人在作案过程中使用的工具。 |

表 1: 命名实体标签及定义

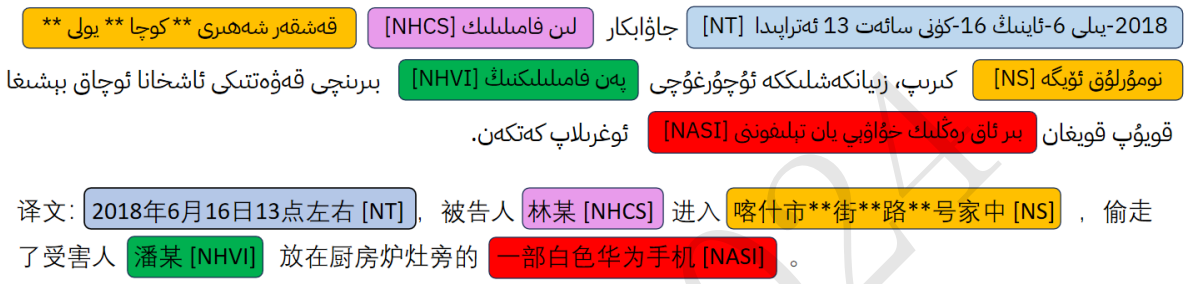



图 1: 标注数据案例

为两组，对同一数据进行两轮标注：第一轮作为预标注，标注完成后对遇到的问题进行整理和分析，根据情况调整和细化标注方案；第二轮则为正式标注，期间使用使用开源的数据标注平台Label-Studio[4]来监控整个过程，确保数据标注的一致性和质量。为了保证标注数据的准确性，本文从数据集中随机抽取了50个样本，并手动检查样本中的实体是否与标签相符。结果如表2所示，匹配率(Matching Acc)表示在人工评估下有多少个实体与其标签相匹配，总的匹配率达到了86.89%。

| | NHCS | NT | NS | NCGV | NHVI | NCSM | NASI | NO | NATS |
|---|---|---|---|---|---|---|---|---|---|
| **样本数量** | 72 | 12 | 53 | 33 | 35 | 25 | 8 | 23 | 6 |
| **正确标注** | 67 | 10 | 48 | 23 | 30 | 21 | 7 | 21 | 5 |
| **匹配率** | 93.06% | 83.33% | 90.57% | 69.70% | 85.71% | 84.00% | 87.50% | 91.30% | 83.33% |

表 2: 标注实体人工评估结果

## 4 模型

### 4.1 问题描述

司法领域命名实体识别任务的具体目标可以描述为：给定低资源语言法律文书的NER训练集$D = \{(x_i, y_i)\}_{i=1}^{N}$，其中$x_i = (w_1, w_2, ..., w_m)$ 表示输入的法律文本序列，$m$ 为序列中的单词总数。$y_i \in \{$司法实体类型集合$\}$ 为相应的实体标签，$N$ 为训练集中的样本总数。该任务的目标是学习一个映射函数$f : D \rightarrow Y \in \{$司法实体类型集合$\}$，以便准确地预测输入文本序列中的法

[4]https://labelstud.io/


(c) 2024 中国中文信息学会计算语言学专业委员会

律命名实体及其类别。在此过程中，本文重点关注如何有效地扩充词汇并利用有限的标注数据来提升模型对低资源语言法律文书的命名实体识别能力。

### 4.2 多语言预训练模型mBERT

**BERT Multilingual Base**(简称为mbert-base)(Pires et al., 2019)是一个由Google 开发的多语言版自然语言处理模型，基于原始的BERT(Bidirectional Encoder Representations from Transformers)(Devlin et al., 2019)架构。此模型在包括维吾尔语在内的104种语言上进行预训练，能够处理并理解多种语言的文本，适用于多语种自然语言处理任务。mbert-base 采用了Transformer 架构，具有12层的双向编码器，隐藏层大小为768，以及12个自注意力头。这种结构使得模型能够捕捉文本中的复杂语境和依赖关系。此外，模型使用了512个位置嵌入来处理不同长度的输入序列。在预训练阶段，mbert-base主要通过两种方式进行训练：Masked Language Model(MLM) 和Next Sentence Prediction(NSP)。MLM任务通过随机遮盖输入中的单词并预测其身份来增强模型对上下文的理解能力。NSP任务则训练模型预测两个句子是否在原始文本中顺序相连，这有助于模型捕获句子级的语义关系。本文基于维吾尔语法律文书构建的微调数据集，因此采用Hugging Face Transformers框架中mbert-base模型,该模型包含约1.1亿参数。本文将mBERT模型微调20个epoch，初始学习率设置为1e-5，权重衰减率设置为0.01，batch size设置为16，500步的warmup进行训练。

### 4.3 词汇扩充

多语言预训练语言模型mBERT无法识别维吾尔语法律文书中的字符。这意味着当模型预测维吾尔语法律文书中的命名实体时，可能会将很多实体标记为未知[UNK]，从而影响模型命名实体识别的性能。

分词及词表构建。维吾尔语从右到左书写，单词之间使用空格隔开。每个维吾尔语单词都可以作为一个特征项。本文对维吾尔语法律文本的分词处理，采用预训练模型内置的WordPiece分词方法。这是一种基于词级别的分词策略，适用于处理像维吾尔语这样的黏着型语言。通过WordPiece方法获得175257个tokens，通过数据清洗，去掉标签符号、去掉重复字词，最终得到包含的13118个tokens的维吾尔语法律文书专用词汇列表。

词表合并和嵌入矩阵扩展。为了整合新词汇，本文采用了一种扩展词嵌入矩阵的方法。mBERT模型原始词汇表包含119547个tokens。分词模型生成的新词汇表与预训练模型的原始词汇表合并时，只增加预训练模型原始词汇表中没有的字词，扩展后mBERT模型的词汇列表大小为131912个token。接下来，为使模型能处理新增词汇，原始的词嵌入矩阵从$V \times D$被扩展至$V' \times D$，其中$V$ 是初始词汇表的大小，$D$ 是嵌入向量的维度，而$V'$ 表示更新后的词汇表大小。此过程中，新的行被添加到矩阵中，为新词汇提供数值化的词向量表示。这些新增词向量以均值0和方差0.02的高斯分布进行初始化，为新词汇提供了一个合适的数值起点，以便在接下来的训练中进行调整，以此改善新词汇在模型中的语义表征:

$$\vec{v}_{new} \sim \mathcal{N}(0, 0.02I) \tag{1}$$

其中，$\vec{v}_{\text{new}}$ 表示新词嵌入的词向量，$\vec{v}_{\text{new}} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ 表示该向量遵循均值为0，方差为$\sigma^2$的正态分布，$\mathbf{I}$ 代表单位矩阵。

嵌入向量调整。为了让初始化的新词汇嵌入向量有效地融入到mBERT的语义空间中，本文利用维吾尔语法律文书的标注数据集UgLaw-NERD进行微调训练，目的是调整新词汇的嵌入向量，使之更准确地反映词汇的语义特性和上下文相关性。嵌入向量调整通过优化以下目标函数来实现:

$$L(\theta) = -\sum_{(x,y)\in D}\sum_{c\in C} y_c \log(f_c(x;\theta)) \tag{2}$$

$L(\theta)$ 是关于模型参数$\theta$ 的损失函数，$D$ 是标注数据集，$x$ 是输入文本，$y$ 是与$x$ 对应的实体标注向量，$C$ 是所有可能的类别集合，$y_c$ 是真实标注中类别$c$ 的指示(0或1)，$f_c(x;\theta)$是mBERT模型预测输入文本$x$ 属于类别$c$ 的概率。这个损失函数计算了对于数据集中每个样本真实标注和模型预测之间的差异总和。


(c) 2024 中国中文信息学会计算语言学专业委员会

## 5 实验与分析

在本节中，从数据集与评价指标、对比实验、消融实验和参数分析四个部分介绍实验过程和细节。

### 5.1 数据集与评价指标

实验采用本文构建的维吾尔法律文书命名实体数据集UgLaw-NERD，本数据集包含9种司法实体类型，共96745个实体。实验中按照8:1:1的比例将数据集划分为训练集、验证集和测试集，数据集详细信息如表3所示：

| 数据集 | NHCS | NT | NS | NCGV | NHVI | NCSM | NASI | NO | NATS |
|---|---|---|---|---|---|---|---|---|---|
| 训练集 | 7458 | 12186 | 25256 | 6149 | 3518 | 2718 | 17589 | 2525 | 650 |
| 验证集 | 863 | 1416 | 2939 | 724 | 417 | 298 | 1984 | 340 | 67 |
| 测试集 | 913 | 1472 | 3194 | 807 | 391 | 398 | 2094 | 288 | 91 |
| **UgLaw-NERD** | 9234 | 15074 | 31389 | 7680 | 4326 | 3414 | 21667 | 3153 | 808 |

表 3: 维吾尔语法律文书命名实体数据集统计

本文用精确率P(Precision)、召回率R(Recall)和综合评价指标F1(F1-Measure)作为实验结果的评价指标，应用序列标注任务的seqeval脚本来评估UgLaw-NERD数据集在模型命名实体识别任务上的性能。

### 5.2 对比实验

本文采用过去研究中主要用于维吾尔语命名实体识别的以下基线模型进行对比实验：

- **CRF**:通过经典CRF模型提取维吾尔语法律文本特征进行命名实体识别。
- **Bi-LSTM**:利用可以更好的捕捉双向语义依赖关系的Bi-LSTM模型。
- **Bi-LSTM+CNN+CRF**:采用Bi-LSTM模型结合CNN和CRF进行命名实体识别。
- **mBERT-base(F.t)**:多语种预训练语言模型mBERT在数据集上微调进行维吾尔语法律文书的命名实体识别。
- **mBERT-base(F.t.voc)**:采用融合词汇扩充的多语种预训练语言模型mBERT在数据集上微调。

| 数据集 | UgLaw-NERD | | |
|---|---|---|---|
| 模型 | P | R | F1 |
| CRF | 74.85% | 72.53% | 73.67% |
| Bi-LSTM | 79.53% | 81.83% | 80.67% |
| Bi-LSTM+CNN+CRF | 80.60% | 81.76% | 81.18% |
| mBERT-base(F.t) | 81.19% | 83.50% | 82.33% |
| mBERT-base(F.t.voc) | **88.96%** | **90.48%** | **89.72%** |

表 4: UgLaw-NERD数据集上实验结果，(F.t)表示在数据集上微调的模型，(F.t.voc)表示融合词汇扩充后在数据集上微调的模型，加粗表示最优实验结果。

实验结果如表4所示，从表中可以得到如下结论：(1)本文提出的融合预训练模型词汇扩充和微调的mBERT-base(F.t.voc)模型在UgLaw-NERD数据集上得到了最好的结果，其F1分数为89.72%，相较于现有的基线模型mBERT-base(F.t)提升了7.39%。这一显著的提升证明了通过词汇扩充进行模型微调的有效性，特别是在处理语言多样性和专业术语较多的维吾尔语法律


(c) 2024 中国中文信息学会计算语言学专业委员会

文书数据集上。(2)从表中可见，mBERT-base模型在与传统CRF和Bi-LSTM模型的比较中表现更佳。这种性能提升部分得益于Transformer架构的全局注意力机制，它优化了上下文信息的捕捉。同时，mBERT的预训练背景也为其性能增强提供了支持。(3)Bi-LSTM+CNN+CRF模型组合相比单独的Bi-LSTM模型在F1分数上提升了0.51%，这表明结合不同类型的神经网络可以提升模型对细节的捕捉能力，特别是在实体边界识别和语义特征提取方面。(4)对于CRF模型，虽然在性能上不及深度学习模型，但其在数据集上的表现仍显示出一定的基线竞争力。综上所述，通过将维吾尔语司法领域的专有词汇添加到mBERT-base模型词汇表中，显著提升了其对维吾尔语中低频词汇和特定法律术语的理解与处理能力。词汇扩充扩大了模型的语言覆盖范围，使得在命名实体识别任务中能够实现更准确的预测。这一改进有效增强了模型在维吾尔语法律文书中识别命名实体的能力。

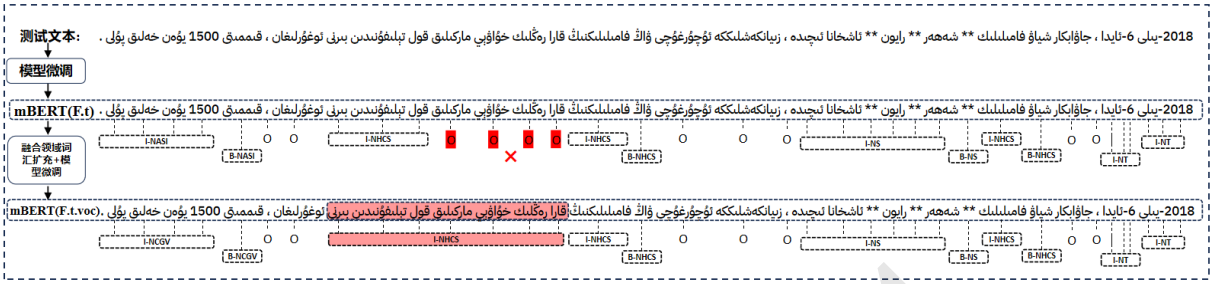



图 2: 样例图

图2给出了维吾尔语司法领域命名实体识别样例。在没有融合领域词汇扩充的情况下，仅对模型进行微调可能无法准确识别所有相关实体。这表明融合领域词汇扩充结合模型微调，能够有效提高维吾尔语法律文书命名实体识别的性能，不仅改善了模型的总体识别精度，还特别增强了对法律术语的识别能力。

### 5.3 消融实验

为了验证融合词汇扩充的预训练模型命名实体识别效果的有效性，本文在UgLaw-NERD数据集上进行消融实验，并设计以下模型变体：-N表示原始模型；(F.t)表示仅在数据集上微调的模型；(F.t.voc)表示融合词汇扩充和在数据集上微调的模型。

| 模型 | P | R | F1 |
|---|---|---|---|
| mBERT-base-N | 0.37% | 1.78% | 0.61% |
| mBERT-base(F.t) | 81.19% | 83.50% | 82.33% |
| mBERT-base(F.t.voc) | **88.96%** | **90.48%** | **89.72%** |

表 5: 消融实验结果，加粗表示最优实验结果。

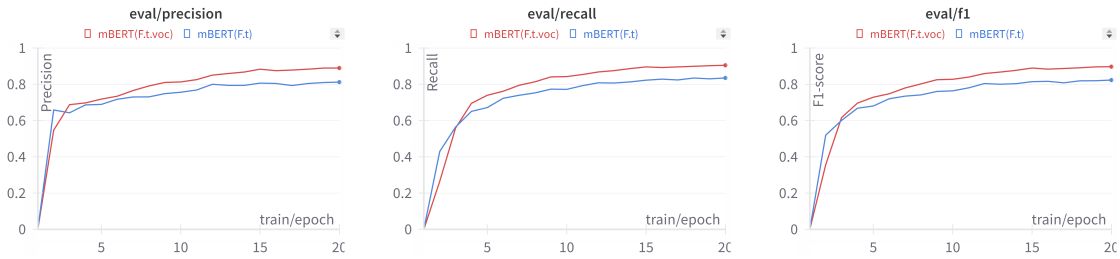



图 3: 模型在训练过程的表现，蓝色曲线代表mBERT-base(F.t)、红色曲线代表mBERT-base(F.t.voc)。

消融实验结果如表5和图3所示。从表和图中可以得到如下结论：(1)mBERT-base-N模型的性能极低，几乎没有实用价值，其精确率为0.37%，召回率为1.78%，F1分数为0.61%。这表明预训练模型直接应用于特定任务而不进行任何调整是不可取的，尤其是在语言特性复杂和语料资源有限的情况下。(2)当mBERT-base模型在UgLaw-NERD数据集上进行微调时，性能显著提升，F1分数达到82.33%。这种提升表明微调对预训练模型适应于低资源语言和特定任务是非常有效的。(3)在mBERT-base模型基础上进行词汇扩充并微调后，模型的性能进一步提升，F1分数提高到89.72%。这表明词汇扩充对于处理维吾尔语这种资源稀缺语言的命名实体识别任务尤其关键。词汇扩充帮助模型更好地理解和处理低资源语言中的低频词汇和复杂表达，从而提高了模型的识别精度和覆盖范围。

### 5.4 参数分析

为了探究词汇量大小对低资源语言法律文书命名实体识别模型性能的影响。本文通过调整模型中词汇量的大小，创建了不同的预训练模型变体，分别为mBERT-0、mBERT-10、mBERT-30、mBERT-50、mBERT-80和mBERT-100，其中数字代表扩充词汇量占本文构建的完整司法领域词汇表中词汇数量的百分比。

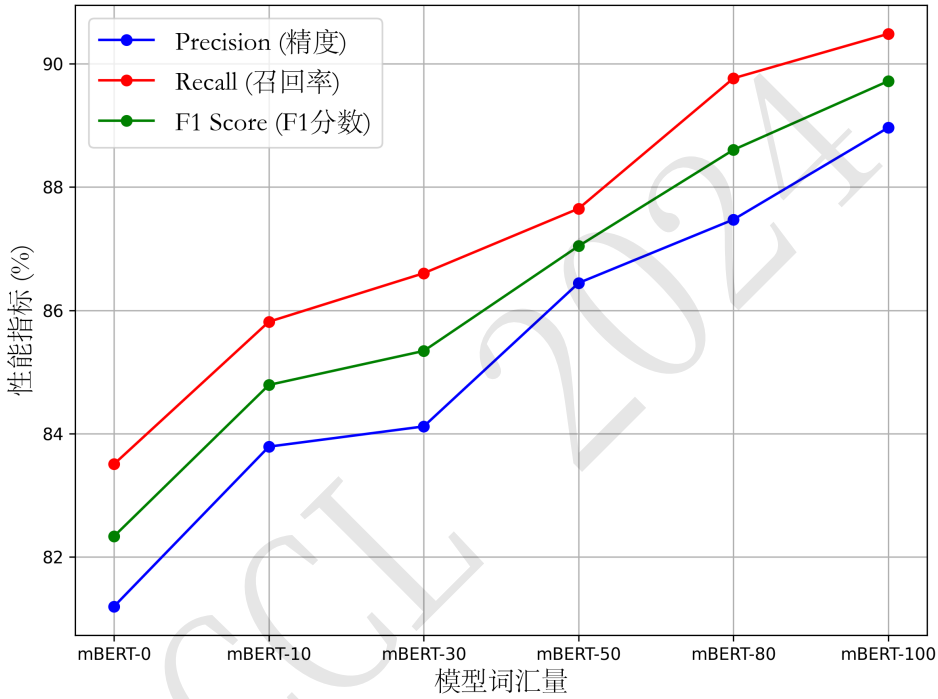



图 4: 新增词汇量大小对模型性能的影响

图4展示了不同词汇量大小对于模型性能的影响。从图中可以看出，词汇扩充量从0%增至100%，模型的精确度、召回率和F1分数均实现了逐步提升。具体来看，mBERT-0在这三个性能指标上的表现是最低的，而mBERT-100则在所有指标上达到了最高值。这一结果表明词汇扩充在增强模型对命名实体识别能力上的直接效果。F1分数作为精确度与召回率的综合指标，其显著的增长证实了词汇扩充策略在改善模型整体性能上的作用。在词汇扩充的各个级别中，性能提升与词汇扩充量呈现线性关系，特别是在词汇扩充量较高的阶段(mBERT-50至mBERT-100)，性能提升较为显著。综上所述，词汇扩充对低资源语言法律文书的命名实体识别模型有显著的正向作用。这表明在处理低资源语言的命名实体识别任务时，增加特定领域的词汇来扩充模型的词汇量是提高性能的一种有效策略。

## 6 总结与展望

针对现有多语言预训练语言模型在低资源语言法律文书命名实体识别任务中应用的局限性，本文提出了一种融合词汇扩充的改进策略。该方法通过收集维吾尔语司法领域词汇，动


(c) 2024 中国中文信息学会计算语言学专业委员会

态扩展了预训练模型的词汇表，并利用人工标注的维吾尔语法律文书命名实体数据集UgLaw-NERD上进行微调，提高了模型对维吾尔语法律文书中命名实体的识别能力。实验结果表明，与现有基准模型相比，本文提出的方法在性能上取得了明显提升。未来，本文将尝试把融合预训练模型词汇扩充的方法应用于更多低资源语言，进一步验证和提高方法的泛化性和实用性，从而为推进低资源语言处理技术的发展做出贡献。

## 参考文献

Kahaerjiang Abiderexiti, Maihemuti Maimaiti, Tuergen Yibulayin, and Aishan Wumaier. 2017. Construction of uyghur named entity relation corpus. *International Journal of Asian Language Processing*, 27(2):155–172.

Azmat Anwar, Xiao Li, Yating Yang, Rui Dong, and Turghun Osman. 2020. Constructing uyghur named entity recognition system using neural machine translation tag projection. In *Chinese Computational Linguistics: 19th China National Conference, CCL 2020, Hainan, China, October 30–November 1, 2020, Proceedings 19*, pages 247–260. Springer.

Luiz Henrique Bonifacio, Paulo Arantes Vilela, Gustavo Rocha Lobato, and Eraldo Rezende Fernandes. 2020. A study on the impact of intradomain finetuning of deep language models for legal named entity recognition in portuguese. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9*, pages 648–662. Springer.

Yanguang Chen, Yuanyuan Sun, Zhihao Yang, and Hongfei Lin. 2020. Joint entity and relation extraction for legal documents with legal feature enhancement. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1561–1571.

Yahui Chen. 2015. Convolutional neural network for sentence classification. Master's thesis, University of Waterloo.

Hyung Won Chung, Dan Garrette, Kiat Chuan Tan, and Jason Riesa. 2020. Improving multilingual models with language-clustered vocabularies. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4536–4546, Online, November. Association for Computational Linguistics.

Michael Collins and Yoram Singer. 1999. Unsupervised models for named entity classification. In *1999 Joint SIGDAT conference on empirical methods in natural language processing and very large corpora*.

Junjie Deng, Hanru Shi, Xinhe Yu, Wugedele Bao, Yuan Sun, and Xiaobing Zhao. 2023. Milmo: minority multilingual pre-trained language model. In *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 329–334. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Hideki Isozaki and Hideto Kazawa. 2002. Efficient support vector classifiers for named entity recognition. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

John Lafferty, Andrew McCallum, Fernando Pereira, et al. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Icml*, volume 1, page 3. Williamstown, MA.

Elena Leitner, Georg Rehm, and Julian Moreno-Schneider. 2019. Fine-grained named entity recognition in legal documents. In *International Conference on Semantic Systems*, pages 272–287. Springer.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.


(c) 2024 中国中文信息学会计算语言学专业委员会

Wen Tai, H. T. Kung, Xin Dong, Marcus Comiter, and Chang-Fu Kuo. 2020. exBERT: Extending pre-trained models with domain-specific vocabulary under constrained training resources. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1433–1439, Online, November. Association for Computational Linguistics.

Hirotaka Tanaka and Hiroyuki Shinnou. 2022. Vocabulary expansion of compound words for domain adaptation of BERT. In Shirley Dita, Arlene Trillanes, and Rochelle Irene Lucas, editors, *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation*, pages 379–387, Manila, Philippines, October. Association for Computational Linguistics.

Henry S Thompson and Alex Lascarides. 1999. Ninth conference of the european chapter of the association for computational linguistics. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*.

Kiyotaka Uchimoto, Qing Ma, Masaki Murata, Hiromi Ozaku, and Hitoshi Isahara. 2000. Named entity extraction based on a maximum entropy model and transformation rules. In *proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 326–335.

Wenwen Xu, Mingzhe Fang, Li Yang, Huaxi Jiang, Geng Liang, and Chun Zuo. 2021. Enabling language representation with knowledge graph and structured semantic information. In *2021 International Conference on Computer Communication and Artificial Intelligence (CCAI)*, pages 91–96. IEEE.

Ziqing Yang, Zihang Xu, Yiming Cui, Baoxin Wang, Min Lin, Dayong Wu, and Zhigang Chen. 2022. CINO: A Chinese minority pre-trained language model. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3937–3949, Gyeongju, Republic of Korea, October. International Committee on Computational Linguistics.

Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. 2014. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*.

买合木提·买买提, 王路路, 吐尔根·依布拉音, 艾山·吾买尔, and 卡哈尔江·阿比的热西提. 2019. 基于条件随机场的维吾尔文机构名识别. 计算机工程与设计, 40:273–278.

加日拉·买买提热衣木, 吐尔根·依布拉音, and 艾山·吾买尔. 2014. 基于统计和规则混合策略的维吾尔人名识别研究. 新疆大学学报*(*自然科学版*)*, 31:319–324.

吾守尔, 斯拉木, 帕丽旦, 杨文忠, et al. 2018. 基于bilstm-cnn-crf 模型的维吾尔文命名实体识别. 计算机工程, 44(8):230–236.

孔祥鹏, 吾守尔, 斯拉木, 杨启萌, and 李哲. 2020. 基于迁移学习的维吾尔语命名实体识别. 东北师大学报*:*自然科学版, 52(2):58–65.

木合塔尔·艾尔肯, 艾斯卡尔·艾木都拉, and 地里木拉提·吐尔逊. 2013. 基于规则的维吾尔地名识别. 通信技术, 46:103–105.

李春楠, 王雷, 孙媛媛, and 林鸿飞. 2021. 基于bert 的盗窃罪法律文书命名实体识别方法. 中文信息学报, 35(8):73–81.

汪昆, 帕力旦, 吐尔逊, et al. 2017. 统计与规则相结合的维吾尔语人名识别方法. 自动化学报, 43(4):653–664.

王路路, 艾山·吾买尔, 买合木提·买买提, 卡哈尔江·阿比的热西提, and 吐尔根·依布拉音. 2018. 基于crf和半监督学习的维吾尔文命名实体识别. 中文信息学报, 32:16–26+33.

王路路, 艾山, 吾买尔, 吐尔根, et al. 2019. 基于深度神经网络的维吾尔文命名实体识别研究. 中文信息学报, 33(3):64–70.

艾斯卡尔·肉孜, 宗成庆, 姑丽加玛丽·麦麦提艾力, 热合木·马合木提, and 艾斯卡尔·艾木都拉. 2013. 基于条件随机场的维吾尔人名识别方法. 清华大学学报*(*自然科学版*)*, 53:873–877.

阿迪来, 冯向萍, et al. 2017. 基于条件随机场的维吾尔语音乐实体识别. 智能计算机与应用, 7(2):59–62.

麦合甫热提, 米日姑·肉孜, 麦热哈巴·艾力, and 吐尔根·依布拉音. 2014. 基于语法语义知识的维吾尔文机构名识别. 计算机工程与设计, 35:2944–2948.


(c) 2024 中国中文信息学会计算语言学专业委员会

# 基于动态提示学习和依存关系的生成式结构化情感分析模型

**贾银涛,崔佳佳,穆玲玲*,咎红英**
郑州大学计算机与人工智能学院, 河南省郑州市, 450001
ytjia2024@gs.zzu.edu.cn, jjcui@gs.zzu.edu.cn, iellmu@zzu.edu.cn, iehyzan@zzu.edu.cn

## 摘要

结构化情感分析旨在从文本中抽取所有由情感持有者、目标事物、观点表示和情感极性构成的情感元组，是较为全面的细粒度情感分析任务。针对目前结构化情感分析方法错误传递，提示模版适应性不足和情感要素构成复杂的问题，本文提出了基于动态提示学习和依存关系的生成式结构化情感分析模型，根据不同的情感元组构成情况分别设计提示模版，并用模板增强生成式预训练模型的输入，用依存关系增强生成效果。实验结果显示，本文提出的模型在SemEval2022[1]数据集上的SF1值优于所对比的基线模型。

**关键词：** 结构化情感分析 ； 提示学习 ； 依存关系

# Dynamic Prompt Learning and Dependency Relation based Generative Structured Sentiment Analysis Model

**Yintao Jia, Jiajia Cui, Lingling Mu*, Hongying Zan**
School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou Henan, 450001
ytjia2024@gs.zzu.edu.cn, jjcui@gs.zzu.edu.cn, iellmu@zzu.edu.cn, iehyzan@zzu.edu.cn

## Abstract

Structured sentiment analysis aims to extract all sentiment tuples composed of holder, target, expression and polarity from the text, which is a more comprehensive fine-grained sentiment analysis task. Aiming at the problems of error propagation, insufficient adaptability of prompt templates and complex composition of sentiment tuples in current structured sentiment analysis methods, this paper proposes a dynamic prompt learning and dependency relations based generative structured sentiment analysis model. Prompt templates are designed according to different composition of sentiment tuples, and templates are used to enhance the input of the generative pre-trained model. Augment generation with dependency relationships. The experimental results show that the SF1 value of the proposed model on the SemEval2022 dataset is better than the compared baseline models.

**Keywords:** Structured sentiment analysis , Prompt learning , Dependency relations

[1]https://competitions.codalab.org/competitions/33556


(c) 2024 中国中文信息学会计算语言学专业委员会

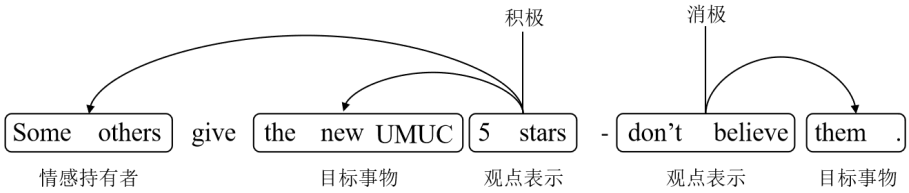



图 1. 结构化情感分析样例

## 1 引言

结构化情感分析(Structured Sentiment Analysis, SSA) (Barnes et al., 2021)旨在抽取文本中所有的情感元组，情感元组由情感持有者(holder, h)、目标事物(target, t)、观点表示(expression, e)和情感极性(polarity, p)构成，目的是分析情感持有者对目标事物表达的情感极性，其中情感极性通过观点表示来表达。结构化情感分析样例如 1所示。相较于方面级情感分析，结构化情感分析从句子中抽取的情感元素更多更复杂，并且针对的是目标对象整体评价，而方面级情感分析针对的是目标对象的方面或属性进行评价。

结构化情感分析在实际应用中具有非常重要的意义，如在舆情分析方面，结构化情感分析能够自动识别各类人群对于热点事件的情感表示和情感倾向，对政府了解民意，及时做出具体调整具有一定的意义；在商品和服务评论上，结构化情感分析能够自动识别各类用户对产品或者产品的某个方面属性所发表的观点及情感倾向，对消费者挑选产品或服务和企业优化产品都有明确的指导和建议作用。

目前在结构化情感分析中主要的研究方法有图分析方法、序列标注方法和生成式方法。 (Barnes et al., 2021)首次提出将结构化情感分析转化为关系图分析任务，该方法将观点表示作为根节点，通过与其他元素连接的弧表示它们之间的关系。 (Zhai et al., 2023)通过改进情感元素标签和情感元素内部词与词的关系标签解决情感元素在句子中不连续和重叠的问题。序列标注方法通常用在管道方法的第一阶段对情感元素进行抽取，然后通过第二阶段中的关系判断模块对第一阶段抽取到的情感元素之间的关系进行判断 (Poświata, 2022; Barikbin, 2022; Anantharaman et al., 2022)。生成式方法主要是通过构建提示模板直接将结构化情感分析任务转化为序列到序列的文本生成任务，目前在结构化情感分析任务中的生成式方法设计的提示模板是构建不连续的文本序列，模型直接生成情感元素位置索引或者带有特殊标签的情感元素文本(Morio et al., 2022; Raghav et al., 2022)。

上述方法虽然在结构化情感分析任务上取得了一定的成绩，但是存在以下问题。（1）图分析方法对关系图的设计有着很高的要求，关系图设计的好坏对模型整体性能有着决定性的影响；（2）序列标注方法存在错误传递的问题；（3）生成式方法中设计的提示模板生成不连续的序列，对于情感元素之间的关系不能充分表示，而且对情感元组构成不同的情况使用静态的提示模板，不利于模型区分不同情况的情感元组。

结构化情感分析任务的数据集主要有$OpeNER_{EN}$、MPQA和$DS_{Unis}$等，数据集以观点表示为基础标注情感元组，其中情感持有者相对于观点表示数量很少，而且目标事物的数量不大于观点表示的数量，即目标事物可能为空。除此之外，$OpeNER_{EN}$大量句子包含多个情感元组，多个数据集中标注的情感元素不连续。因此，数据集中存在的问题是（1）完整的情感元组很少；（2）情感元组的构成情况复杂。

针对上述研究方法的局限性和数据集的问题，本文提出了基于动态提示学习和依存关系的生成式结构化情感分析模型(dynamic prompt learning and dependency syntactic relations based generative structured sentiment analysis model, $DPLG_{Dep}$)。在模型中针对不同情感元组动态使用提示模板，并且将依存关系融入到生成式预训练模型中提高模型性能。本文的研究工作如下：

- 提出了根据不同情感元组构成设计动态提示模版的思想，针对完整的情感四元组和情感持

*通讯作者
**项目基金：**期货领域黑嘴检测研究，横向项目，郑州易盛信息技术有限公司。
©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

有者为空的情感元组分别设计提示模板将其转化为连续的自然语句形式。

- 提出了用提示模板增强原始输入的方法对预训练模型微调，增强预训练模型在下游任务上的知识。
- 提出了用依存关系与预训练模型融合的方法提高模型对情感元素之间的关系建模，实验结果表明融入依存关系之后，模型性能有所提升，对于依存关系复杂的情感元素的识别能力也得到增强。

## 2 相关工作

**提示学习** 提示学习是将原始输入语句$x$通过使用提示转化为一个提示字符串$x'$，这个提示字符串中有一些空的位置，使用预训练语言模型利用概率去补充这些空置的位置，从而得到最终的输出$y$ (Liu et al., 2023)。在提示学习中，提示模板的构建是基础的工作，也是本文研究方法中的重要研究内容。提示模板的形式主要包括完形填空式和前缀式，完形填空式提示模板会在提示模板中空出一部分位置，这些位置需要使用预训练模型预测的结果进行填充，这种形式的提示模板常用在信息抽取任务中，如对“Readers think the book is boring”进行情感分析，设计提示“The book is [MASK]”，其中[MASK]就是模型要预测的内容;而前缀式提示模板则是在输入句子和目标答案之间设计一些词或者特殊符号进行提示，在文本生成任务如机器翻译和摘要生成任务中常用，如机器翻译中设计提示模板为“English: [X] Chinese: [Y]”，其中[X]是输入的英文句子，[Y]是模型生成的中文的翻译文本。在本文针对不同的情感元组，设计了对应的完形填空式的提示模板，然后为增强提示效果，将提示模板与输入句子进行拼接作后缀。

**生成式情感分析方法** 生成式情感分析方法通过设计提示模板，将文本中包含的情感元素填充到提示模板对应位置形成目标序列，然后利用生成式预训练模型将情感分析任务转化为文本生成任务。模型输出的目标序列有两种形式：一种是不连续的文本序列形式，如直接生成情感元素位置索引或者带有特殊标签的情感元素文本，另一种是连续的自然语言句子形式的目标序列。不连续的文本序列形式在第1节中已经介绍，下面介绍自然语言句子为目标序列的生成式情感分析方法。 (Liu et al., 2021)在方面类型情感分析任务（ACSA）任务中设计自然语言形式的提示模板，将任务转换为自然语言生成任务。 (Zhang et al., 2021b)提出直接给生成式预训练模型输入句子中的情感元素添加注释，将经过注释的输入句子作为模型生成的目标句子。 (Zhang et al., 2021a)在方面情感四元组预测任务中设计自然语言句子形式的提示模板，不仅可以直接识别四种情感元素，而且也能够充分利用自然语言形式的标签中的语义信息。然而这些生成式方法使用的提示模板为静态模板，本文在 (Zhang et al., 2021a)的基础上提出动态提示模板。

**融入句法信息的情感分析方法** 在预训练模型的基础上添加句法或语义信息的知识对于解决结构化情感分析任务是有效的，能够增强预训练模型获取文本句子中结构化信息的能力，提升模型的性能 (Sarangi et al., 2022; Yu et al., 2023)。 (Sarangi et al., 2022)在结构化情感分析任务中，使用BERT (Devlin et al., 2018)模型获取输入句子的上下文表示，同时使用分析器和神经网络获取句子的语义特征，最后将上下文表示和语义特征进行拼接，这种融合的特征比只使用BERT获取的上下文表示的效果更好。 (Yu et al., 2023)在方面类型情感分析任务中首次提出将依存关系融入到生成式预训练模型中，使用图注意网络 (Veličković et al., 2018)获取依存特征表示，同时使用BART (Lewis et al., 2020)的编码器获取输入句子的上下文表示，然后将两种特征采用门控机制进行融合。在本文中同样使用门控机制将依存关系与预训练模型T5 (Raffel et al., 2020)进行融合。

## 3 模型

本文设计的$DPLG_{Dep}$模型包括两个模块：基于动态提示学习的生成模块和句法分析模块。在生成模块中根据输入文本句子包含的情感元组使用对应的提示模板，在句法分析模块获取输入文本句子的依存特征，然后使用融合方法将依存特征融合到生成式预训练模型中。

### 3.1 基于动态提示学习的生成模块

生成模块的主体是一个编码器-解码器架构的生成式预训练模型T5，输入序列包括两部分，一部分是原始的输入文本句子，另一部分则是输入文本句子包含的情感元组对应的提示模板。目标序列是情感元素填充到提示模板中对应的位置得到的文本序列。如图 2中输入文本句


(c) 2024 中国中文信息学会计算语言学专业委员会

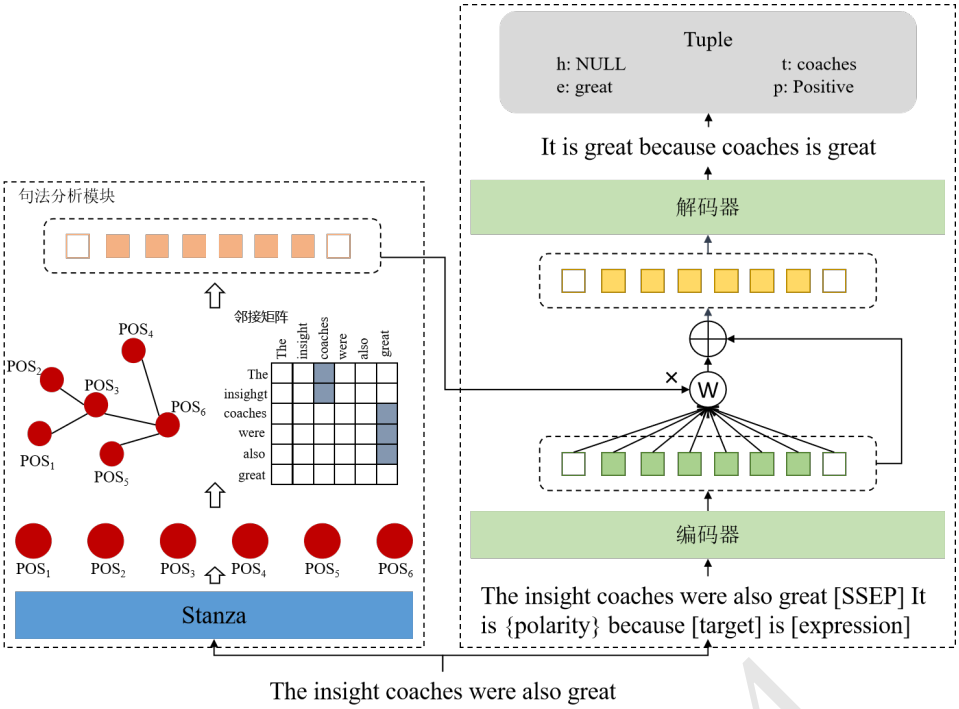



图 2. $DPLG_{Dep}$模型

子“The insight coaches were also great”包含情感元组(NULL, “coaches”, “great”, “Positive”)，根据这个情感元组得到对应的提示模板，将输入句子和提示模板使用“[SSEP]”拼接，得到输入序列。然后将情感元素填充到提示模板中对应的位置得到模型输出的目标序列“It is great because coaches is great”。

本文提出根据不同情感元组构成设计动态提示模版的思想，所以当情感元组中四个情感元素都存在时，设计的提示模板为：

[h] said [t] is [e], [p]

其中[h]表示情感持有者，[t]表示目标事物，[e]表示观点表示，[p]表示情感极性对应的词。当情感元组中情感持有者为空时，设计对应的提示模板为：

It is [p] because [t] is [e]

当情感元组为空时，使用空字符串$\emptyset$作为提示模板。

本文对目标事物进行判断，如果目标事物为空则用“it”代替，否则将目标事物填充到模板中：

$$[t] = \begin{cases} [t] & [t] \neq \emptyset \\ it & [t] = \emptyset \end{cases} \tag{1}$$

情感极性分为积极(Positive)，中性(Neutral)和消极(Negative)，直接将情感极性填充到提示模板中不符合自然语言的表达，而且在不同的提示模板中，情感极性在句子中充当的语义角色也不同，所以本文需要将情感极性映射为可以表达对应情感的词，并且针对不同的情况需要转化成不同的词。当情感元组中情感持有者存在时，在对应的提示模板中情感极性的映射过程为：

$$[p] = \begin{cases} satisfactorily, & Positive \\ objectively, & Neutral \\ dissaprovingly, & Negative \end{cases} \tag{2}$$

当情感持有者为空时，对应的提示模板中情感极性的映射过程为：

$$[p] = \begin{cases} great, & Positive \\ ok, & Neutral \\ bad, & Negative \end{cases} \tag{3}$$


(c) 2024 中国中文信息学会计算语言学专业委员会

如果输入语句中包括多个情感元组，那么给每个情感元组生成对应的目标语句$y_i$，并使用“ [SSEP] ”将目标语句进行拼接作为预训练的生成模型的输出即$Target = y_1[SSEP]y_2[SSEP]\ldots y_k$。

针对数据集中存在的情感元素的短语不连续的问题，本文直接将不连续的短语使用特殊符号“ [SEP] ”进行拼接，以区分在标准答案中情感元素连续和不连续的情况，优点是更符合标准答案形式，减少在模型训练过程添加额外的噪音。由情感元组线性化为目标句子以及对于不连续情感元素处理的样例如表 1所示。其中输入句子1中包含两种情感元组，第一个情感元组为完整情感元素四元组，第二个为情感持有者为空的情感元组。输入句子2中情感持有者和目标事物都为空。在输入句子3中展示了标注不连续的观点表示[“definitely going”，“again”]，在转化为目标句子时，转化为“definitely going [SEP] again”。

| 文本类型 | 文本序列 |
|---|---|
| 输入句子1 | **Clean sizeable room** with **good facilities** in the part of the hotel **we** were in . |
| 情感元组 | h: we, t: facilities, e: good, p: Positive |
| 目标句子 | we said facilities is good, satisfactorily |
| 情感元组 | h:NULL, t: room, e: Clean sizeable, p: Positive |
| 目标句子 | It is great because room is Clean sizeable |
| 输入句子2 | **Bit pricey** and but away from center |
| 情感元组 | h: NULL, t: NULL , e: Bit pricey, p: Negative |
| 目标句子 | It is bad because it is Bit pricey |
| 输入句子3 | **I**’m **definitely going there again** whenever I get a chance . |
| 情感元组 | h: I, t: there, e: [**definitely going**, **again**], p: Positive |
| 目标句子 | I said there is **definitely going [SEP] again**, satisfactorily |

表 1. 目标句子构建和非连续情感元素处理

本文考虑只将提示模板经过情感元素填充之后形成目标句子，在这个过程中预训练模型学习到的结构化情感分析的信息不够充分，并且比较难理解情感元素在目标句子中的语义角色，因此，将提示模板添加到输入句子的后面作为后缀，给预训练模型提示更多的信息。

### 3.2 句法分析模块

**依存关系计算**：本文使用邻接矩阵表示依存关系，对输入句子使用句法分析工具Stanza (Qi et al., 2020)进行分析，获取到句子的句法信息包括词性，(依存关系，头结点，尾结点)的依存关系三元组，然后构建依存关系的邻接矩阵$A \in R^{n\times n}$，构造过程如公式4所示，需要注意的是，为符合结构化情感分析任务特征，本文的依存关系保留词在句子中的“root”关系，如图 3，“hotel”就是句子的依存关系的根节点，使用邻接矩阵中的(0，index)表示，其中index是“hotel”在句子中的位置索引，在图 3中index的值为7：

$$A_{i,j} = \begin{cases} 1, & \text{词i,j之间有依存关系} \\ 0, & \text{词i,j之间无依存关系} \end{cases} \tag{4}$$

**节点特征表示**：边节点特征使用分析得到的词性表示，根据分析器得到的句子的词性$P = [pos_1, pos_2, pos_3, \ldots, pos_n]$，其中$pos_1$前添加一个代表“root”节点对应的词性，如图 3，本文使用“$< pad >$”作为“root”的词性。然后初始化一个$Embedding_{pos}$，获取词性的特征表示。

$$E^{pos} = Embedding_{pos}(P) \tag{5}$$

其中$E^{pos} \in R^{n\times d}$，$n$是预训练模型T5设置的最大长度，$d$表示模型T5的维度。

**依存特征的获取**：本文使用图注意网络GAT获取依存特征表示。使用词性嵌入作为GAT模型的节点特征，使用依存关系的邻接矩阵$A$实现掩码注意力。GAT模型获取依存特征表示的计算：

$$h_{l+1} = ELU(a_l h_l W_l) \tag{6}$$


(c) 2024 中国中文信息学会计算语言学专业委员会

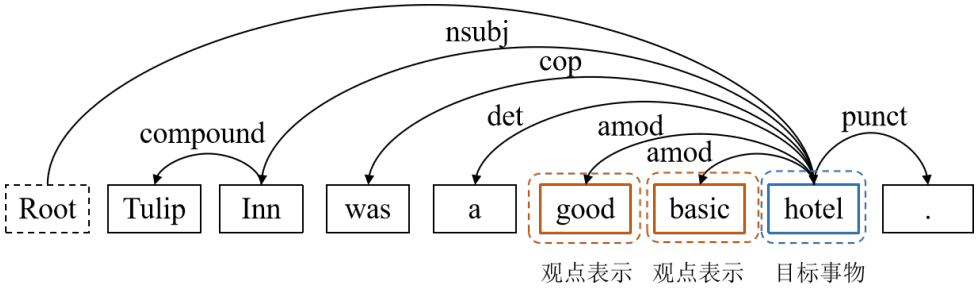



图 3. 依存分析样例

$$a_l = Dropout(soft\max(e_l)) \tag{7}$$

$$e_l = Leaky\mathrm{Relu}(h_l W_l a[:d,:] + (h_l W_l a[d:,:])^T) \tag{8}$$

其中$ELU$为激活函数，$h_l \in R^{n\times d}$为第$l$层句子的词性特征表示，$W \in R^{d\times d}$，$a \in R^{2d\times 1}$，$h_0 = E^{pos}$，$l$设置为2。将GAT模型最后一层隐状态表示作为依存特征表示$E^{dep}$。

预训练模型T5模型编码器的输出包括对输入句子$x$和提示模板$Template$拼接之后的序列的输出，所以编码器输出的语义特征$O_{encoder}$为：

$$O_{encoder} = Encoder_{T5}(x + [SSEP] + Template) \tag{9}$$

本文为了能够使得获取的依存特征与T5模型的编码器的输出$O_{encoder}$在形式上尽可能的一致，在构建依存关系的邻接矩阵时对其边长进行填充和裁剪，使得邻接矩阵的形状固定，填充采用的T5中tokenizer的“$< pad >$”特殊字符，并且在$E^{dep}$开始和结尾添加T5模型中表示序列开始和结束的特殊符号的嵌入，在T5-base模型中，开始符号默认为“$< pad >$”，结束符号默认为“$< /s >$”。所以$E^{dep}$最终还需进行转换：

$$E^{dep} = Embedding_{T5}(< pad >)||E^{dep}[:-2,:]||Embedding_{T5}(< /s >) \tag{10}$$

### 3.3 特征融合方式

本文在将依存关系特征表示$E^{dep}$与T5的输出$O_{encoder}$进行融合时设计了四种融合方式，包括直接相加，拼接，门控机制 (Yu et al., 2023)和交叉注意力。

**直接相加**：由于在$E^{dep}$已经将开始和结束位置替换为T5的开始和结束位置，所以直接相加在理论上可行的，计算方式为：

$$H = O_{encoder} + E^{dep} \tag{11}$$

**拼接**：直接对$O_{encoder}$和$E^{dep}$进行拼接，然后经过全连接层，对拼接之后的表示进行维度的转换以及获取拼接之后更重要的信息。

$$H = [O_{encoder}||E^{dep}]W^{trans} + b^{trans} \tag{12}$$

**门控机制**：这种方法能够根据编码器输出的上下文表示动态的调整依存特征的权重，具体就是使用一个线性网络对$O_{encoder}$进行维度转换，然后使用激活函数获取到权重值，对依存特征加权，然后与$O_{encoder}$做残差连接。

$$H = O_{encoder} + W^O E^{dep} \tag{13}$$

其中$W \in R^{d\times 1}$。

**交叉注意力**：本文考虑在将依存特征$E^{dep}$与编码器输出的上下文表示$O_{encoder}$进行融合时，$E^{dep}$和$O_{encoder}$的权重都应该是动态调整的，但是经过交叉注意力之后输出的特征表示与$O_{encoder}$存在较大差距，所以需要额外使用一个上下文表示进行求和，做残差连接，所以在交叉注意力中，将$O_{encoder}$作为查询，$E^{dep}$作为键和值。

$$H = O_{encoder} + Att * E^{dep} \tag{14}$$

$$Att = softmax(\frac{O_{encoder} * (E^{dep})^T}{\sqrt{d}}) \tag{15}$$


(c) 2024 中国中文信息学会计算语言学专业委员会

### 3.4 训练目标

将融合之后的特征用$e$表示，而解码器之前生成的结果用$y_{<i}$表示，解码器解码当前输出的概率分布$y_i$需要基于$e$和$y_{<i}$进行计算：

$$y_i = Decoder_{T5}(e, y_{<i}) \tag{16}$$

模型训练时采用最大似然估计，对T5模型的输入序列和输出序列进行映射训练更新两个模块的模型参数：

$$\max_{\theta} \log f_\theta(y|e) = \sum_{i=1}^{n} \log f_\theta(y_i|e, y_{<i}) \tag{17}$$

## 4 实验设置

### 4.1 数据集

本文在$OpeNER_{EN}$ (Agerri et al., 2013)、MPQA (Wiebe et al., 2005)和$DSUnis$ (Toprak et al., 2010)数据集上进行实验。三个数据集的统计信息如表 2所示。

| 数据集 | | 句子数 | 情感持有者 | 目标事物 | 观点表示 | 情感极性 |
|---|---|---|---|---|---|---|
| $DS_{Unis}$ | 训练集 | 2253 | 63 | 806 | 806 | 806 |
| | 验证集 | 232 | 9 | 98 | 98 | 98 |
| | 测试集 | 318 | 12 | 130 | 130 | 130 |
| MPQA | 训练集 | 5873 | 1425 | 1481 | 1706 | 1706 |
| | 验证集 | 2063 | 406 | 494 | 570 | 570 |
| | 测试集 | 2112 | 434 | 462 | 518 | 518 |
| $OpeNER_{EN}$ | 训练集 | 1744 | 266 | 2679 | 2884 | 2884 |
| | 验证集 | 249 | 49 | 371 | 400 | 400 |
| | 测试集 | 499 | 98 | 793 | 865 | 865 |

表 2. 数据集统计信息

### 4.2 参数设置

$DPLG_{Dep}$使用的生成式预训练模型为T5-base，本文训练时batch size设置为16，epoch设置为20，学习率设置为3e-4，设置的输入句子的最大长度为128，模型维度为768，预测时采用贪心解码策略。在句法分析模块实现中，包括邻接矩阵的构建、节点特征的设置、依存特征的捕获以及特征融合方式的设置，模型在NVIDIA RTX3090显卡上加速训练,参数信息如表 3所示。

| 参数类型 | 参数值 |
|---|---|
| 预训练模型 | T5-base |
| 输入最大长度 | 128 |
| 预训练模型维度 | 768 |
| 邻接矩阵大小 | 128×128 |
| 节点特征维度 | 768 |
| GAT层数 | 2 |
| GAT的dropout | 0.5 |
| 交叉注意力头数 | 8 |

表 3. 参数设置

CCL 2024

### 4.3 基线模型

与本文的模型$DPLG_{Dep}$进行对比的模型有图分析模型，序列标注模型和生成式模型。

**图分析模型**：Graph-parser (Barnes et al., 2021)是结构化情感分析任务的基线模型，USSA (Zhai et al., 2023)在Graph-parser的基础上改进以解决数据集中情感元素不连续问题。

**序列标注模型**：SeqL (Anantharaman et al., 2022)仅通过双层BiLSTM (Graves and Graves, 2012)模型实现序列标注和情感元素关系判断的模型，TBST (Poświata, 2022)首先使用基于Transformer (Vaswani et al., 2017)的模型实现序列标注，然后使用R-BERT [1]模型判断情感元素之间的关系。

**生成式模型**：GenA (Raghav et al., 2022)直接使用BART模型从输入句子生成情感元素索引，SeqGen (Morio et al., 2022)使用T5-large模型从输入句子生成带有标签的情感元素，以及直接使用大语言模型Llama3 [2]预测的方法。

### 4.4 评价指标

在结构化情感分析任务中，为了评估模型预测的情感元组的性能，提出使用情感图F1(Sentiment Graph F1, **SF1**) (Samuel et al., 2022)作为评估指标，在SF1的计算中，一个正确的样例就是情感元素完全匹配，否则计算预测的情感元素与标准答案的情感元素之间重叠的词数权重，然后取情感持有者、目标事物和观点表示这三个情感元素权重的平均值作为当前情感元组的权重值。

## 5 结果及分析

### 5.1 模型结果

| 模型 | 类型 | 数据集 | | |
|---|---|---|---|---|
| | | $\mathbf{OpeNER_{EN}}$ | **MPQA** | $\mathbf{DS_{Unis}}$ |
| Graph-parser | 图分析方法 | 52.1 | 12.5 | 20.4 |
| USSA | | - | 30.5 | 33.2 |
| SeqL | 序列标注方法 | 32.9 | 0.02 | 0.06 |
| TBST | | 67.0 | 32.6 | 39.5 |
| GenA | 生成式方法 | 62.6 | 32.7 | 33.0 |
| SeqGen | | <u>67.1</u> | <u>40.9</u> | <u>40.5</u> |
| Llama3 | | 18.3 | 0.1 | 4.9 |
| $DPLG_{Dep}$ | | **70.4** | **43.3** | **51.6** |

表 4. $DPLG_{Dep}$模型结果(SF1, %)

表 4中展示了$DPLG_{Dep}$模型与其他方法的对比结果，$DPLG_{Dep}$模型的结果是在5个随机种子的平均值。从表中可以看到，本文提出的模型在所有数据集上的效果达到了最佳，并且生成式方法相较于图分析方法和序列标注方法都有更好的性能。在图分析方法中可以看到不同的关系图的设计会导致模型的性能会有很大的差距，其中USSA方法在解决数据集中情感元素不连续和重叠的问题，相较于Graph-parser方法提升效果明显。在序列标注方法中，可以看到仅仅使用BiLSTM作为节点抽取和关系判断在结构化情感分析任务中的效果不理想，而通过使用预训练模型可以极大地提高模型的性能。而在生成式方法中，相较于GenA和SeqGen模型中使用静态的提示模板生成不连续的目标序列，本文提出的$DPLG_{Dep}$模型效果更好，证明使用动态提示模板，并且用提示模板对原始句增强可以使得T5模型学习到更多的结构化情感分析的知识，而SeqGen方法中使用的模型为T5-large，相较于本文方法使用的T5-base，耗费的资源更多。而没有微调直接使用Llama3进行预测的效果比较差。

第二十三届中国计算语言学大会论文集，第442页–第453页，太原，中国，2024年7月25日至28日。
卷1：主会论文
(c) 2024 中国中文信息学会计算语言学专业委员会

| 模型 | 特征融合方式 | 数据集 | | | 平均值 |
|---|---|---|---|---|---|
| | | $\mathbf{OpeNER_{EN}}$ | **MPQA** | $\mathbf{DS_{Unis}}$ | |
| DPLG | - | 69.6 | 42.1 | **52.2** | 54.6 |
| $DPLG_{Dep}$ | 直接相加 | 0.2 | 0.0 | 0.0 | 0.1 |
| | 拼接 | 2.9 | 0.4 | 1.3 | 1.5 |
| | 交叉注意力 | 69.8 | 43.0 | 49.3 | 54.0 |
| | 门控机制 | **70.4** | **43.3** | 51.6 | **55.1** |

表 5. 特征融合方式结果(SF1, %)

### 5.2 特征融合方式结果

本文针对将依存关系特征与编码器输出的上下文表示进行融合的方式也进行了对比验证，实验结果如表 5所示。其中DPLG模型表示$DPLG_{Dep}$消去句法分析模块的模型。由表可知，当使用门控机制时模型的性能最好，说明使用门控机制将依存特征融合到T5-base模型时对模型整体框架产生的影响最小。使用直接相加和拼接方式时模型的性能剧烈下降，在T5模型中已经包含丰富的语义信息，依存信息中可能包含与其重叠或者冲突的信息，所以在直接进行拼接或者相加时，模型需要处理这些冗余或者冲突，可能导致信息混淆。

### 5.3 消融实验结果

| 模型 | 数据集 | | | 平均值 |
|---|---|---|---|---|
| | $\mathbf{OpeNER_{EN}}$ | **MPQA** | $\mathbf{DS_{Unis}}$ | |
| DPLG-suffix | 65.0 | 33.0 | 39.3 | 45.8 |
| DPLG | 69.6 | 42.1 | **52.2** | 54.6 |
| $DPLG_{Dep}$ | **70.4** | **43.3** | 51.6 | **55.1** |

表 6. 消融实验(SF1, %)

本文对$DPLG_{Dep}$模型进行了消融实验，验证各个模块的作用，其中DPLG-suffix模型表示在$DPLG_{Dep}$的基础上，不加依存特征，同时在生成式模型的输入序列中不加提示模板作为后缀。从表中可以看到，在将提示模板添加到输入句子中时，在三个数据集上的性能都有很高的提升，说明将提示模板拼接到生成式预训练模型的输入序列中可以提示模型捕获结构化情感分析任务的信息，帮助模型理解各个情感元素在目标句子中的语义角色。而在DPLG的基础上融入依存关系特征对模型性能有所提升，证明仅仅通过T5模型的注意力识别情感元素的性能有待提高，并且这也是生成式预训练模型的预训练目标与结构化情感分析任务的差异，而融入句法依存关系能够减小这种差异，提高模型对情感元素之间关系的建模能力。

### 5.4 案例分析

在本节对$DPLG_{Dep}$预测的结果与DPLG预测的结果进行对比，如表 7所示，$DPLG_{Dep}$模型相较于DPLG，可以对情感元素根据依存关系进行纠正以及帮助预训练模型捕获模型注意力忽略的句法关系。在样例1中，使用DPLG没有捕获目标事物“it”，而添加依存信息之后，能够捕获到“really great”所修饰的“it”元素。在样例2中，DPLG模型已经准确的预测到第一个情感元组，但是预测第二个情感元组时模型的注意力更多的集中在“everything”上，而将“had a great time !”作为修饰语，根据依存信息，其中关系更强的是“great”对于“time”的修饰，所以$DPLG_{Dep}$可以准确的预测到两个情感元组。在样例3中，DPLG模型本身强大的性能已经可以准确的预测到其中的三个情感元组，$DPLG_{Dep}$中的句法分析模块能够帮助预训练模型捕获“Clean”与“Hotel”的关系。综上可知，在DPLG的基础上添加句法分析模块，能够利用依存信息对预训练模型的注意力捕获到一些存在偏差的情感元素进行纠正，还可以使用依存关系对模型忽略的情感元组关系进行补充，帮助模型对情感元素之间的关系进行建模。

[1]https://github.com/monologg/R-BERT
[2]https://github.com/meta-llama/llama3


(c) 2024 中国中文信息学会计算语言学专业委员会

| 样例来源 | 样例文本 |
|---|---|
| 标准样例1 | For people who want some rest and relaxatio **it** is **really great** . [(NULL, it, really great, Positive)] |
| DPLG | For people who want some rest and relaxatio it is **really great** . [(NULL, NULL, really great, Positive)] |
| $DPLG_{Dep}$ | For people who want some rest and relaxatio **it** is **really great** . [(NULL, it, really great, Positive)] |
| 标准样例2 | **everything** was **quite good** , had a **great time** ! [(NULL, everything, quite good, Positive), (NULL, time, great, Positive)] |
| DPLG | **everything** was **quite good** ,**had a great time** ! [(NULL, everything, quite good, Positive), (NULL, everything, had a great time, Positive)] |
| $DPLG_{Dep}$ | **everything** was **quite good** , had a **great time** ! [(NULL, everything, quite good, Positive), (NULL, time, great, Positive)] |
| 标准样例3 | **Wonderful Hotel** , **Clean Friendly very spacious** [(NULL, Hotel, Wonderful, Positive), (NULL, Hotel, Clean, Positive), (NULL, Hotel, Friendly, Positive) , (NULL, Hotel, very spacious, Positive)] |
| DPLG | **Wonderful Hotel** , Clean **Friendly very spacious** [(NULL, Hotel, Wonderful, Positive), (NULL, Hotel, Clean, Positive), (NULL, Hotel, Friendly, Positive) , (NULL, Hotel, very spacious, Positive)] |
| $DPLG_{Dep}$ | **Wonderful Hotel** , **Clean Friendly very spacious** [(NULL, Hotel, Wonderful, Positive), (NULL, Hotel, Clean, Positive), (NULL, Hotel, Friendly, Positive) , (NULL, Hotel, very spacious, Positive)] |

表 7. $DPLG_{Dep}$与DPLG预测案例对比

### 5.5 模型泛化性验证

为验证本文设计的模型的泛化性，额外增加了数据集进行实验，但是包含情感持有者、目标事物、观点表示和情感极性的英文数据集只有$OpeNER_{EN}$、MPQA和$DS_{Unis}$，而这些数据集有一个共同特征，就是大多数情感持有者为空，所以本文在包含目标事物、观点表示和情感极性的ASTE-Data-V2数据集 (Xu et al., 2020)上进行额外的实验，对模板泛化性进行验证。

**基线模型：**与$DPLG_{Dep}$对比的基线模型有Paraphrase(Zhang et al., 2021a)、BART-ABSA(Yan et al., 2021)、Span-BiDir(Chen et al., 2022)和CONTRASTE(Mukherjee et al., 2023)。

| 模型 | 14Lap | 14Res | 15Res | 16Res |
|---|---|---|---|---|
| BART-ABSA | 58.7 | 65.3 | 59.3 | 67.6 |
| Paraphrase | 60.4 | 71.5 | 62.1 | 71.9 |
| Span-BiDir | 62.7 | **74.3** | 64.8 | 72.1 |
| CONTRASTE | 62.9 | 74.0 | **66.1** | **74.2** |
| $DPLG_{Dep}$ | **66.2** | 73.1 | 64.8 | 73.9 |

表 8. 模型在ASTE-Data-V2数据集的结果对比（F1，%）

从表8中可以看到，相较于当前在ASTE-Data-V2数据集上的先进模型，$DPLG_{Dep}$也表现出具有竞争性的性能，所以本文的模型具有一定的泛化能力。

## 6 总结与展望

针对当前结构化情感分析任务中研究方法的局限性和数据集的问题，本文提出了基于动态提示学习和依存关系的生成式结构化情感分析模型$DPLG_{Dep}$，在生成式模块中根据情感元组的构成情况动态地设计提示模板，并将提示模板添加到输入句子的后面作为后缀给预训练模型T5提示更多的信息。为了减小生成式预训练模型T5的预训练目标与结构化情感分析


(c) 2024 中国中文信息学会计算语言学专业委员会

任务之间的差异和增强模型建模情感元素之间关系的能力，在生成式模块外添加句法分析模块，将句法分析模块获取的依存特征融合到预训练模型中，模型性能得到提升。实验结果显示，$DPLG_{Dep}$模型的性能优于所有对比的基线模型，并且通过案例分析发现，添加句法分析模块，能够利用依存信息对预训练模型的注意力捕获到一些存在偏差的情感元素进行纠正，还可以使用依存关系对模型忽略的情感元组关系进行补充。

## 参考文献


Rodrigo Agerri, Montse Cuadros, Sean Gaines, and German Rigau. 2013. Opener: Open polarity enhanced named entity recognition. *Procesamiento del Lenguaje Natural*, (51):215–218.

Karun Anantharaman, K Divyasri, Jayannthan Pt, S Angel, Rajalakshmi Sivanaiah, Sakaya Milton Rajendram, and TT Mirnalinee. 2022. Ssn_mlrg1 at semeval-2022 task 10: Structured sentiment analysis using 2-layer bilstm. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1324–1328.

Sadrodin Barikbin. 2022. Slpl-sentiment at semeval-2022 task 10: Making use of pre-trained model's attention values in structured sentiment analysis. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1382–1388.

Jeremy Barnes, Robin Kurtz, Stephan Oepen, Lilja Øvrelid, and Erik Velldal. 2021. Structured sentiment analysis as dependency graph parsing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3387–3402.

Yuqi Chen, Chen Keming, Xian Sun, and Zequn Zhang. 2022. A span-level bidirectional network for aspect sentiment triplet extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4300–4309.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Alex Graves and Alex Graves. 2012. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, pages 37–45.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Jian Liu, Zhiyang Teng, Leyang Cui, Hanmeng Liu, and Yue Zhang. 2021. Solving aspect category sentiment analysis as a text generation task. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4406–4416.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Gaku Morio, Hiroaki Ozaki, Atsuki Yamaguchi, and Yasuhiro Sogawa. 2022. Hitachi at semeval-2022 task 10: Comparing graph-and seq2seq-based models highlights difficulty in structured sentiment analysis. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1349–1359.

Rajdeep Mukherjee, Nithish Kannen, Saurabh Pandey, and Pawan Goyal. 2023. Contraste: Supervised contrastive pre-training with aspect-based prompts for aspect sentiment triplet extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12065–12080.

Rafał Poświata. 2022. Opi at semeval-2022 task 10: Transformer-based sequence tagging with relation classification for structured sentiment analysis. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1366–1372.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108.



(c) 2024 中国中文信息学会计算语言学专业委员会

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

R Raghav, Adarsh Vemali, and Rajdeep Mukherjee. 2022. Etms@ iitkgp at semeval-2022 task 10: Structured sentiment analysis using a generative approach. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1373–1381.

David Samuel, Jeremy Barnes, Robin Kurtz, Stephan Oepen, Lilja Øvrelid, and Erik Velldal. 2022. Direct parsing to sentiment graphs. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 470–478.

Pratyush Sarangi, Shamika Ganesan, Piyush Arora, and Salil Joshi. 2022. Amex ai labs at semeval-2022 task 10: contextualized fine-tuning of bert for structured sentiment analysis. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1296–1304.

Cigdem Toprak, Niklas Jakob, and Iryna Gurevych. 2010. Sentence and expression level annotation of opinions in user-generated discourse. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 575–584.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *International Conference on Learning Representations*.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39:165–210.

Lu Xu, Hao Li, Wei Lu, and Lidong Bing. 2020. Position-aware tagging for aspect sentiment triplet extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2339–2349.

Hang Yan, Junqi Dai, Tuo Ji, Xipeng Qiu, and Zheng Zhang. 2021. A unified generative framework for aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2416–2429.

Chengze Yu, Taiqiang Wu, Jiayi Li, Xingyu Bai, and Yujiu Yang. 2023. Syngen: A syntactic plug-and-play module for generative aspect-based sentiment analysis. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Zepeng Zhai, Hao Chen, Ruifan Li, and Xiaojie Wang. 2023. Ussa: A unified table filling scheme for structured sentiment analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14340–14353.

Wenxuan Zhang, Yang Deng, Xin Li, Yifei Yuan, Lidong Bing, and Wai Lam. 2021a. Aspect sentiment quad prediction as paraphrase generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9209–9219.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021b. Towards generative aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 504–510.

CCL 2024

# 基于方面引导的图文渐进融合的多模态方面级情感分析方法

闫自达[1,2]，郭军军[*1,2]，余正涛[1,2]
1.昆明理工大学，信息工程与自动化学院，昆明，650500
2.昆明理工大学，云南省人工智能重点实验室，昆明，650500
2641668065@qq.com, guojjgb@163.com
ztyu@hotmail.com

## 摘要

多模态方面级情感分析旨在通过结合图像信息和文本信息来识别特定方面的情感极性。然而，图像和文本作为两种不同的模态，其在数据表现形式和语义表达上存在显著差异，缩小模态鸿沟和跨模态特征融合是多模态方面级情感分析任务中出现的两个关键问题。对此，本文提出了一种基于方面引导的图文渐进融合的多模态方面级情感分析方法，该方法采用图像和文本中重叠的方面信息作为枢轴，利用方面引导的图文对比学习和基于对比的跨模态语义交互来缩小模态差异、促进语义交互，然后在多模态特征空间中整合视觉和文本信息，通过方面引导的基于对比的多模态语义融合来促进跨模态特征融合，从而提升多模态情感分析的性能。在三个多模态方面级情感分析基准数据集上的实验结果证明了本文提出方法的有效性，并且优于其他大多数最先进的多模态方面级情感分析方法。

**关键词：** 情感分析；语义交互；对比学习；多模态融合

# Aspect-Guided Progressive Fusion of Text and Image for Multimodal Aspect-Based Sentiment Analysis

**Zida Yan[1,2], Junjun Guo[*1,2], Zhengtao Yu[1,2]**
1. Faculty of Information Engineering and Automation, Kunming University of Science and Technology Kunming 650500, China
2. Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology Kunming 650500, China
2641668065@qq.com, guojjgb@163.com
ztyu@hotmail.com

## Abstract

Multimodal Aspect-Based Sentiment Analysis aims to identify the sentiment polarity of specific aspects by combining image and text information. However, images and text are two types modalities, exhibit significant differences in data representation and semantic expression. Narrowing the modal gap and cross-modal feature fusion are two key challenges in multimodal aspect-based sentiment analysis. To address these issues, this paper proposes a aspect-guided progressive fusion of text and image for multimodal aspect-based sentiment analysis method, which uses the overlapping aspect information in visual and textual modalities as a pivot, utilizes aspect-guided contrastive cross-modal semantic interaction to narrow the modal differences, and then integrates visual and textual information in the multimodal feature space. It promotes cross-modal feature fusion through aspect-guided contrastive multimodal semantic fusion, thereby

*郭军军（通讯作者）：guojjgb@163.com
**基金项目**：国家自然科学基金（62162037, 62266027, U21B2027, 62266028）；云南省基础研究计划项目（202301AT070444,202001AT070047, 202001AT070046）


(c) 2024 中国中文信息学会计算语言学专业委员会

enhancing the performance of multimodal sentiment analysis. Experimental results on three multimodal aspect-based sentiment analysis benchmark datasets have proven the effectiveness of the proposed method, which outperforms most other state-of-the-art multimodal aspect-based sentiment analysis methods.

**Keywords:** Sentiment analysis , Semantic interaction , Contrastive learning , Multimodal fusion

## 1 引言

多模态方面级情感分析（*Multi-modal Aspect-Based Sentiment Analysis*, MABSA）的目标是将视觉信息融入文本中，以准确地识别句子中特定方面的情感极性。这一领域近年来获得了国内外研究者的广泛关注，相较于传统的仅基于文本的方面级情感分析，通过整合图像信息，能显著提升情感分析的性能。这项技术在社交媒体、医疗保健、教育等多个领域均展现出广泛的应用前景。

在多模态方面级情感分析研究中，图文两种模态间存在明显的语义差距，跨模态的表示学习和语义对齐通常较难。例如，如图1所示，文本描述了结婚的事件，而图像则展示了一位悲伤的男子。文本和图像之间存在显著的语义鸿沟，只有少数重叠的语义枢纽，如实体“Jennifer”和“Ross”以及视觉表达如“cryin”，才能用来弥合这两种模态之间的语义鸿沟。




图 1. 多模态方面级情感分析任务的示例。给定一张图片、一条推文和特定方面目标，我们的目标旨在预测每个方面的情感极性。

最近的许多研究通过采用几种面向方面的多模态融合策略来应对MABSA中的视觉-文本的对齐挑战，如基于方面的跨模态注意力 (Yu and Jiang, 2019; Yu et al., 2020; Zhang et al., 2021)、视觉-文本相关的多任务学习(Yu et al., 2022b)、视觉-文本跨模态预训练(Yu et al., 2022a)等。然而，这些方法中的大多数倾向于面向方面的特征级融合，直接将视觉、文本和方面表示作为输入，通常忽略了不同模态之间的表示差距。

跨模态语义交互和多模态特征融合是多模态融合过程中必须考虑的两个相互关联的问题。通过在模态空间中进行跨模态语义交互对齐，可以更容易地在多模态特征空间中融合它们的表示。对齐和融合是两个相互依存的子任务，对于成功的多模态融合都是必不可少的。此外，方面通常是图像和文本的重叠语义信息，可作为促进模态语义交互和特征融合的合适枢纽。

为了解决MABSA任务中出现的模态对齐难融合难的问题，本文提出了一种基于方面引导的图文渐进融合的多模态方面级情感分析方法，采用方面为枢纽，引入基于方面的三层对比学习，旨在逐层细化模态对齐，以层次化地促进图文特征融合。提出方面引导的图文对比学习来缩小模态差异；提出方面引导的基于对比的跨模态语义交互模块，通过双层跨模态注意力机制来促进语义交互；提出方面引导的基于对比的多模态语义融合模块，通过对称的mix_up层(Zhang et al., 2018)和多模态融合层来促进跨模态特征融合。通过这种层次化的对齐和融合策略，我们能够更有效地整合图文信息，显著提高面向方面的情感极性识别的准确性。我们的主要贡献如下：

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

- 本文提出基于方面引导的图文渐进融合的多模态情感分析方法，该方法通过渐进式地对齐和融合图文表示，获得了基于方面的多模态表示，从而提高模型对特定方面情感辨别能力。
- 本文提出了基于对比的跨模态语义交互模块，利用方面作为语义支点，指导图像与文本进行跨模态语义交互，以缩小模态间隙。为了促进跨模态特征融合，我们提出了基于对比的多模态语义融合模块，通过利用对称的双层跨模态交互和对比学习，我们实现了多模态语义的融合。
- 在三个基准数据集上对MABSA任务的实验结果证明了我们提出的模型的有效性和鲁棒性，取得了有希望的结果。

## 2 相关工作

多模态方面级的情感分析（*Multi-modal Aspect-Based Sentiment Analysis*, MABSA）通过整合图像和文本信息，旨在确定句子中特定方面的情感极性。MABSA领域在自然语言处理（*Natural Language Processing*, NLP）中受到日益增长的关注，研究了多种方法以应对图文方面级情感分析的挑战。早期方法着重于运用注意力机制以捕捉多模态数据之间的相关性。Xu et al. (2019)提出一个交互记忆力网络，旨在捕获多模态数据中的复杂关联。Yu and Jiang (2019)提出了TomBERT预训练语言模型，通过引入特定的模态注意力和跨模态交互模块以增强情感分析的性能。Zhou et al. (2021)采用基于多模态交互层的对抗训练策略，将文本和视觉表示对齐至共享空间以用于MABSA任务。然而，图像与文本之间的固有模态差异使多模态融合经常面临模态匹配噪声的挑战。为应对视文融合中的抗噪声问题，Yu et al. (2020)引入了基于双线性交互的门控模块，以缩减模态间的差距，针对实体敏感的多模态情感分类问题。Zhang et al. (2021)提出两个基于记忆的模块，通过捕获内模态与跨模态特征以解决模态对齐问题。进一步，Yu et al. (2022a)引入一个辅助重构模块，弥合文本与视觉表示间的语义鸿沟，实现了情感分析的最新进展。Khan and Fu (2021)采用基于Transformer的图像标题生成模型，通过增强文本表示并减少模态间的语义鸿沟来提升情感识别性能。Zhao et al. (2022)提出了一个知识增强框架（KEF），利用图像中的形容词-名词对来增强MABSA任务中情感分析性能。Yang et al. (2022)提出了一个面向面部敏感表情到情感文本（FITE）方法，使用面部表情作为情绪线索，增强跨模态分析中的情识别能力。Jia et al. (2023)提出一种情感区域识别与融合网络，利用相关情感区域来促进于视觉与文本的多模态融合。Wang et al. (2024)提出了一种全局-局部特征融合的共同注意力，以分层的方式整合全局-局部多模态特征以提高情感分析性能。

## 3 方法

本文提出了一种基于方面引导的图文渐进融合的多模态情感分析方法，模型总体架构如图2所示。它主要包含四个网络：图文编码器、基于对比的跨模态语义交互模块、基于对比的多模态语义融合模块及输出模块。

### 3.1 任务描述

给定一组多模态样本集$M$，每个样本$c \in M$ 由一句文本$s$、一张与文本相关的图像$v$、一个方面目标$t$ 以及一个情感标签$y$ 组成，即$c = \{\mathrm{s}, \mathrm{v}, \mathrm{t}, \mathrm{y}\}$。其中文本$s = (w_1, w_2, w_3, \ldots, w_m)$，由$m$ 个单词组成，而方面目标$t = \{t_1, t_2, \ldots, t_n\}$，由$n$ 个单词组成。情感标签$y$ 分为三类：负面、中性和正面，即$y \in \{\mathrm{negative}, \mathrm{neutral}, \mathrm{positive}\}$。目标是学习一个基于方面的情感分类器，以便它能正确预测未见测试集中的方面的情感极性。

### 3.2 图文编码器

#### 3.2.1 文本编码器

文本和方面目标的编码器采用预训练的BERT(Devlin et al., 2019)模型。我们分别在文本$s$和方面术语$t$ 的开始和结束处添加特殊标记[CLS] 和[SEP]，然后将它们输入到预训练的BERT编码器中，以获取文本表示$H_s$ 和方面表示$H_t$，可以表示为，

$$H_s = \mathrm{BERT}(s) \tag{1}$$


(c) 2024 中国中文信息学会计算语言学专业委员会

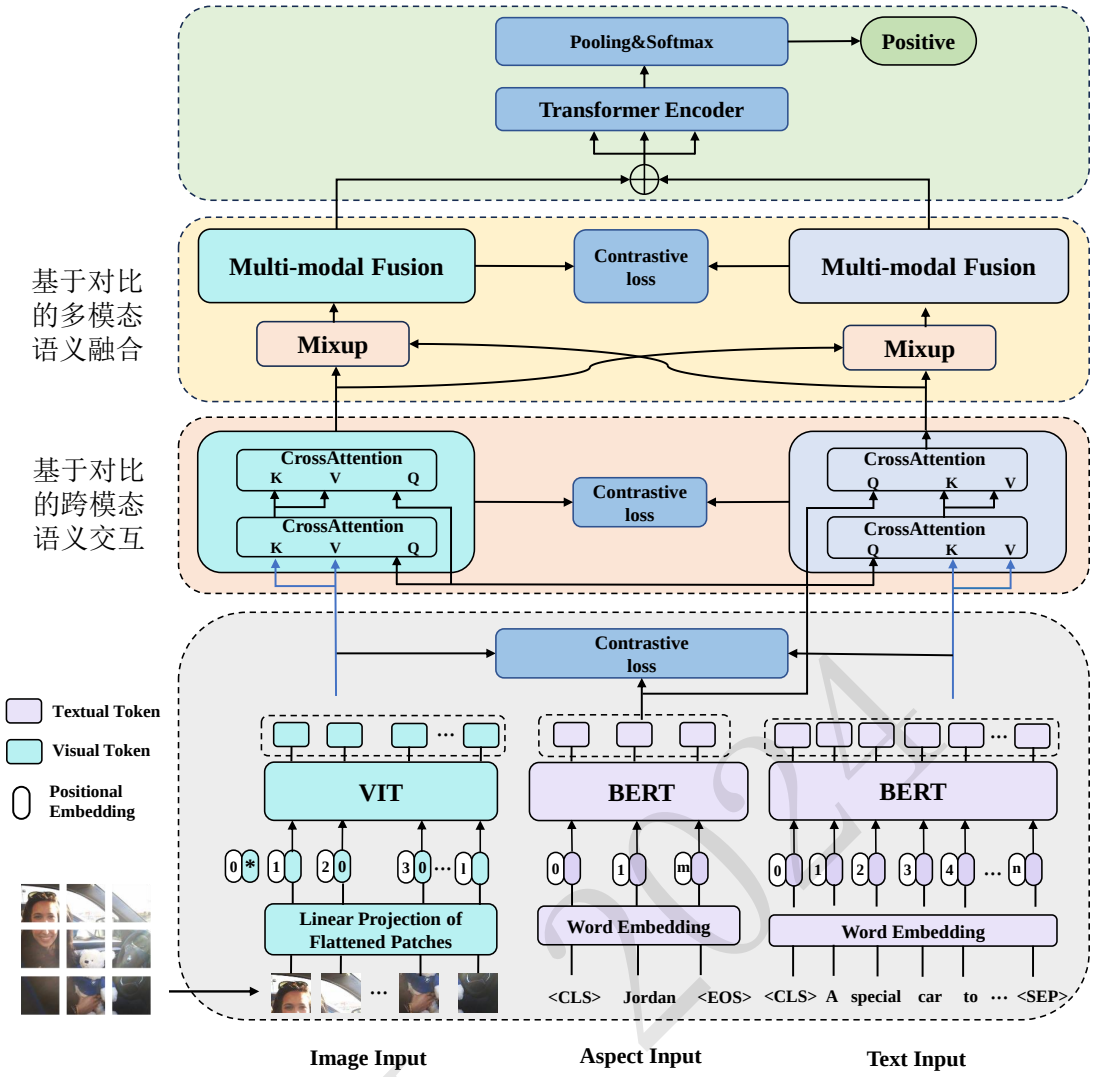



图 2. 基于方面引导的图文渐进融合的多模态情感分析模型

$$H_t = \text{BERT}(t) \tag{2}$$

其中，$H_s \in \mathbb{R}^{m \times d}$, $H_t \in \mathbb{R}^{n \times d}$，$d$ 是隐藏维数，$m$ 和$n$ 分别是输入长度。

### 3.2.2 图像编码器

图像由预训练的Vision Transformer模型（ViT）(Dosovitskiy et al., 2021)来提取初始特征，首先将图像$v$ 切分成$l$ 个图像块序列$[v_1, v_1, \ldots, v_l] \in \mathbb{R}^{l \times (P^2 \times C)}$，然后在前面加上一个特殊的标记$v_{[\text{CLS}]}$，通过线性投影层$V$，再加上位置嵌入得到图像嵌入$\bar{v}$，可以表示为，

$$\bar{v} = \left[v_{[\text{CLS}]}V, v_1V, v_1V, \ldots, v_lV\right] + V^{\text{pos}} \tag{3}$$

其中，$\bar{v} \in \mathbb{R}^{(l+1) \times d}$，线性投影层$V \in \mathbb{R}^{(P^2 \times C) \times d}$，$V^{\text{pos}} \in \mathbb{R}^{(l+1) \times d}$ 是位置嵌入，$d$ 是维数，$(P, P)$ 是每个图像块的分辨率，$C$ 是图像的通道数。进一步地，我们将$\bar{v}$ 输入到Transformer 中，以获取图像表示$H_v \in \mathbb{R}^{(l+1) \times d}$:

$$H_v = \text{Transformer}(\bar{v}) \tag{4}$$

### 3.2.3 方面引导的图文对比学习

为了拟合不同模态的初始表示差距，本文采用方面引导的图文对比学习，以方面表征为枢纽，通过在公共语义空间中拉近图像-方面与文本-方面的距离来促进图文语义对齐，具体可表


(c) 2024 中国中文信息学会计算语言学专业委员会

示如下，

$$L_a = L(H_t, H_s) + L(H_t, H_v) \tag{5}$$

其中，$L_a$ 表示方面引导的图文对比损失，$L(H_t, H_s)$ 表示方面表征和文本表征之间的对比损失，$L(H_t, H_v)$ 表示方面表征和图像表征之间的对比损失，定义如下，

$$L(\alpha, \beta) = \frac{1}{2}(L_{\alpha 2\beta} + L_{\beta 2\alpha}) \tag{6}$$

其中$\alpha, \beta \in \{H_t, H_v, H_s\}$. $L_{\alpha 2\beta}$ 和$L_{\beta 2\alpha}$具体表示如下：

$$L_{\alpha 2\beta} = -\log \frac{\exp(\alpha_i^\top \beta_i/\tau)}{\sum_{j=1}^{N} \exp(\alpha_i^\top \beta_j/\tau) + \sum_{j=1}^{N} \exp(\alpha_i^\top \alpha_j/\tau)} \tag{7}$$

$$L_{\beta 2\alpha} = -\log \frac{\exp(\beta_i^\top \alpha_i/\tau)}{\sum_{j=1}^{N} \exp(\beta_i^\top \alpha_j/\tau) + \sum_{j=1}^{N} \exp(\beta_i^\top \beta_j/\tau)} \tag{8}$$

其中，$\alpha_i$ 和$\beta_j$ 分别是$\alpha$ 中第$i$ 个特征和$\beta$ 中第$j$ 个特征的归一化表示。$N$ 是批次大小，$\tau$ 是温度参数。

### 3.3 基于对比的跨模态语义交互模块

本模块旨在利用方面表示作为跨模态语义支点，引导文本与图像之间的整体语义交互对齐，进而促进跨模态融合。为达成此目的，本文采用了双层跨模态注意力机制来处理方面-文本与方面-图像间的语义交互。首先，以方面表示$H_t$ 作为查询（Query），图像表示$H_v$ 和文本表示$H_s$ 分别作为键（Key）/值（Value），通过添加残差和层归一化操作(Ba et al., 2016) 获得基于方面的浅层图-文表示$H'_v$ 和$H'_s$，如下：

$$H'_v = \mathrm{LN}(H_t + \mathrm{Cross\_Att}(H_t, H_v, H_v)) \tag{9}$$

$$H'_s = \mathrm{LN}(H_t + \mathrm{Cross\_Att}(H_t, H_s, H_s)) \tag{10}$$

其中，$H'_v$, $H'_s \in \mathbb{R}^{n\times d}$，LN表示层归一化操作。

进一步地，继续将$H_t$ 视为查询，将$H'_v$、$H'_s$ 视为键和值，再次通过跨模态注意力机制进行交互，并加入残差及层归一化操作，以获得更深层的基于方面的图-文表示$H''_v$、$H''_s$，如下所示：

$$H''_v = \mathrm{LN}(H_t + \mathrm{Cross\_Att}(H_t, H'_v, H'_v)) \tag{11}$$

$$H''_s = \mathrm{LN}(H_t + \mathrm{Cross\_Att}(H_t, H'_s, H'_s)) \tag{12}$$

其中，$H''_v$, $H''_s \in \mathbb{R}^{n\times d}$。接着，对基于方面的图-文表示$H''_v$ 和$H''_s$ 进行对比学习，以进一步促进文本与图像之间的整体语义对齐。对比学习的损失函数表示如下，

$$L_b = L(H''_v, H''_s) = \frac{1}{2}(L_{v2s} + L_{s2v}) \tag{13}$$

其中，$L_{v2s}$ 和$L_{s2v}$ 分别表示从图像到文本和从文本到图像的对比损失，定义如下：

$$L_{v2s} = -\log \frac{\exp(\alpha_i^\top \beta_i/\tau)}{\sum_{j=1}^{N} \exp(\alpha_i^\top \beta_j/\tau) + \sum_{j=1}^{N} \exp(\alpha_i^\top \alpha_j/\tau)} \tag{14}$$

$$L_{s2v} = -\log \frac{\exp(\beta_i^\top \alpha_i/\tau)}{\sum_{j=1}^{N} \exp(\beta_i^\top \alpha_j/\tau) + \sum_{j=1}^{N} \exp(\beta_i^\top \beta_j/\tau)} \tag{15}$$

其中，$\alpha$ 和$\beta$ 分别代表$H''_v$ 和$H''_s$。


(c) 2024 中国中文信息学会计算语言学专业委员会

### 3.4 基于对比的多模态语义融合模块

为了进一步实现多模态语义的融合，我们引入了一个基于对比的多模态融合模块。首先，通过采用mix_up技术进行基于方面的图文模态的平滑过渡，从而得到浅层多模态表示，以文本融合图像的表示$H_{s2v}$ 和以图像融合文本的表示$H_{v2s}$，具体定义如下：

$$H_{s2v} = w_{\text{mix}} H_s'' + (1 - w_{\text{mix}}) H_v'' \tag{16}$$

$$H_{v2s} = w_{\text{mix}} H_v'' + (1 - w_{\text{mix}}) H_s'' \tag{17}$$

其中，$w_{\text{mix}}$ 为融合权重，$H_{s2v}$ 和$H_{v2s} \in \mathbb{R}^{n \times d}$ 代表浅层的多模态表示。

为避免过度融合而丧失图文特有的信息，我们采用并行的带有归一化的多头自注意力层和多头跨模态注意力层，以保留浅层的多模态表示。具体地，将平滑过渡的$H_{s2v}$ 分别输入自注意力层，同时将$H_{s2v}$ 作为查询，$H_{v2s}$ 作为键和值输入跨模态注意力层，通过集成这两者的输出得到最终基于文本的多模态表示$E_t \in \mathbb{R}^{n \times d}$：

$$E_t = \text{LN}\left[(H_{s2v} + \text{Self\_ATT}(H_{s2v})) + (H_{s2v} + \text{Cross\_ATT}(H_{s2v}, H_{v2s}))\right] \tag{18}$$

其中，Self_ATT 是多头自注意力层，Cross_ATT 是多头跨模态注意力层。

类似于$E_t$，我们将$H_{v2s}$ 输入到自注意力层，将$H_{v2s}$ 视为查询，同时将$H_{s2v}$ 视为键和值输入跨模态注意力层，以此来获取基于图像的多模态表示$E_v \in \mathbb{R}^{n \times d}$：

$$E_v = \text{LN}\left[(H_{v2s} + \text{Self\_ATT}(H_{v2s})) + (H_{v2s} + \text{Cross\_ATT}(H_{v2s}, H_{s2v}))\right] \tag{19}$$

接着，对基于文本的多模态表示$E_t$ 和基于图像的多模态表示$E_v$ 进行多模态对比学习，作为辅助损失函数以进一步促进多模态信息的对齐与融合。损失函数具体表示如下：

$$L_c = L(E_t, E_v) = \frac{1}{2}(L_{v2t} + L_{t2v}) \tag{20}$$

其中，$L_{v2t}$ 和$L_{t2v}$ 分别表示图像到文本和文本到图像的对比损失，定义如下：

$$L_{v2t} = -\log \frac{\exp(\alpha_i^\top \beta_i / \tau)}{\sum_{j=1}^{N} \exp(\alpha_i^\top \beta_j / \tau) + \sum_{j=1}^{N} \exp(\alpha_i^\top \alpha_j / \tau)} \tag{21}$$

$$L_{t2v} = -\log \frac{\exp(\beta_i^\top \alpha_i / \tau)}{\sum_{j=1}^{N} \exp(\beta_i^\top \alpha_j / \tau) + \sum_{j=1}^{N} \exp(\beta_i^\top \beta_j / \tau)} \tag{22}$$

其中，$\alpha$ 和$\beta$ 分别代表$E_t$ 和$E_v$。

### 3.5 输出模块

将基于文本的多模态表示$E_t$ 和基于图像的多模态表示$E_v$ 通过最后一维进行拼接，得到多模态表示$E_M = E_t \,\|\, E_v$，并将其输入到一个Transformer编码器中，我们得到：

$$H = \text{Transformer}(E_M, E_M, E_M) \tag{23}$$

其中，$H \in \mathbb{R}^{n \times 2d}$ 是Transformer编码器输出的最终多模态表示，$\|$ 表示拼接操作。

然后，将池化后的$H$ 送入softmax层进行情感分类：

$$P(y \mid H) = \text{softmax}(\text{Pooling}(H)W) \tag{24}$$

其中，$W \in \mathbb{R}^{d \times 3}$，Pooling 表示池化操作。

使用交叉熵损失作为分类损失$L$ 来学习$W$：

$$L = \frac{1}{N} \sum_{j=1}^{N} P(y^j \mid H) \tag{25}$$

最终的目标函数定义为组合目标函数：

$$L_{\text{all}} = L + \lambda_a L_a + \lambda_b L_b + \lambda_c L_c \tag{26}$$

其中，$\lambda_a$, $\lambda_b$ 和$\lambda_c$ 是超参数。


(c) 2024 中国中文信息学会计算语言学专业委员会

# 4 实验

## 4.1 数据集

我们在三个基准数据集上进行实验，包括Twitter-15 和Twitter-17 [-1]以及MASAD[0] 数据集。Twitter-15 和Twitter-17 数据集由Yu and Jiang (2019)提出，包含了2014-2015 年和2016-2017 年的多模态推文，这些推文覆盖了诸如政治、体育、娱乐等多种主题。每个多模态推文包括一段文本、文本旁的图片、文本中提及的方面术语，以及每个方面的情感标签。情感标签从负面、中性、正面的集合中分配。MASAD 数据集由Zhou et al. (2021) 提出。该数据集涵盖了七个领域内的57 个预定义方面，包括食品、商品、建筑、动物、人类、植物和景观等。方面的情感标签仅包括正面和负面。表1 和表2 提供了这三个基准数据集的详细信息。

| | TWITTER-15 | | | TWITTER-17 | | |
|---|---|---|---|---|---|---|
| Label | Train | Dev | Test | Train | Dev | Test |
| Positive | 928 | 303 | 317 | 1508 | 515 | 493 |
| Neutral | 1883 | 670 | 607 | 1638 | 517 | 573 |
| Negative | 368 | 149 | 113 | 416 | 144 | 168 |
| Total | 3179 | 1122 | 1037 | 3562 | 1176 | 1234 |

表 1. TWITTER-15和TWITTER-17数据集的详细数据

| | Train | | | Test | | | Total | | |
|---|---|---|---|---|---|---|---|---|---|
| | Positive | Negative | Total | Positive | Negative | Total | Positive | Negative | Total |
| Food | 2360 | 433 | 2793 | 592 | 109 | 701 | 2952 | 542 | 3494 |
| Goods | 2671 | 1674 | 4345 | 743 | 512 | 1255 | 3414 | 2186 | 5600 |
| Buildings | 1450 | 970 | 2420 | 367 | 245 | 612 | 1817 | 1215 | 3032 |
| Animal | 3023 | 2208 | 5231 | 1126 | 670 | 1796 | 4149 | 2878 | 7027 |
| Human | 1999 | 1838 | 3837 | 503 | 464 | 967 | 2502 | 2302 | 4804 |
| Plant | 2819 | 2607 | 5426 | 1269 | 947 | 2216 | 4088 | 3554 | 7642 |
| Scenery | 3600 | 1936 | 5536 | 907 | 490 | 1397 | 4507 | 2426 | 6933 |
| Total | 17922 | 11666 | 29588 | 5507 | 3437 | 8944 | 23429 | 15103 | 38532 |

表 2. MASAD数据集的详细数据

## 4.2 参数设置和评价指标

本文采用预训练的`bert-base-uncased` 模型来提取文本和方面的初始特征，同时使用预训练的`vit_base_patch16_224` 模型作为图像编码器。这两种模型的隐藏维度均设置为768，多头自注意力和交叉注意力机制中的头数设置为8，前馈神经网络的内层维度设置为2048。为了防止模型过拟合，我们冻结了`vit` 的参数。模型的所有可训练参数使用AdamW 优化器进行更新。对于数据集Twitter-15 和Twitter-17，学习率分别设置为$1 \times 10^{-5}$ 和$1 \times 10^{-4}$，数据集MASAD学习率设置为$1 \times 10^{-5}$，批次大小均为32。超参数$w_{\text{mix}}$为0.85，$\lambda_a$, $\lambda_b$ 和$\lambda_c$均为0.001。实验在单个RTX 3090 GPU 上进行。

本文采用准确率（Acc）和F1分数（Macro-F1）作为评估指标来衡量本文提出的模型在多模态方面级情感分析任务中的性能。

## 4.3 基线模型

我们采用三种类型的基线模型来评估方面级情感分析的性能，包括仅图像的方法、仅文本的方法和多模态方法，具体如下：

[-1]https://drive.google.com/file/d/1PpvvncnQkgDNeBMKVgG2zFYuRhbL873g/view
[0]https://drive.google.com/file/d/19YJ8vEYCb-uEKUqSGFmysUTvNzxhVKFE/view


(c) 2024 中国中文信息学会计算语言学专业委员会

**仅图像方法：** Res-Aspect(He et al., 2016)：采用Resnet和Bert模型分别提取视觉和方面特征。

**仅文本方法：** MGAN(Fan et al., 2018)：采用多粒度注意力网络，促进文本与方面间的交互进行情感分类。BERT(Devlin et al., 2019)：预训练的Bert模型，通过输入方面目标和文本进行情感分类。BART(Lewis et al., 2020)：一种序列到序列的预训练BART模型，通过输入句子及其方面目标预测情感极性。ATAE-LSTM(Wang et al., 2016)：结合方面信息并通过注意力机制识别句子显著部分的模型。IAN(Ma et al., 2017b)：设计交互注意力机制以映射方面与相应句子间关系的模型。RAM(Chen et al., 2017)：引入加权记忆机制以精确捕捉指定方面相关情绪信息的模型。TNet(Li et al., 2018)：利用CNN层从Bi-RNN派生的转换词表示中提取关键特征的模型。

**多模态方法：** Res-MGAN：设计跨模态多粒度注意力网络，促进文本与图像交互以进行情感分类。Res-Bert：使用预训练的BERT和Resnet模型提取文本和视觉特征，并应用跨模态注意力机制进行多模态融合以进行情感分类。MIMN(Xu et al., 2019)：提出多跳记忆网络捕捉文本与视觉模态间交互的模型。ESAFN(Yu et al., 2020)：提出实体敏感的多模态注意力融合网络，动态捕获方面-文本与方面-图像表示的模型。ViLBERT(Lu et al., 2019)：基于图像与文本训练的多Transformer模型ViLBERT。TomBERT(Yu and Jiang, 2019)：提出面向目标的多模态BERT架构，捕获方面感知的多模态表示，增强多模态情感分类性能。CapTrBERT(Khan and Fu, 2021)：使用BERT通过图像标题和句子作为输入判断情感极性。MMAP(Zhou et al., 2021)：采用对抗训练与多模态层合并文本和图像进行情感分析。saliencyBERT(Wang et al., 2021)：使用递归注意力网络捕获方面敏感的视觉特征。HIMT(Yu et al., 2022a)：使用辅助重构模块的层次交互模型弥合文本和视觉表示间的语义差距。KEF(Zhao et al., 2022)：采用知识增强框架（KEF）从图像中提取形容词-名词对，改善视觉注意力与情感预测。FITE(Yang et al., 2022)：一种应用面部指示文本情绪（FITE）方法，利用面部表情增强跨模态情感识别。ARFN(Jia et al., 2023)：一种情感区域识别与融合网络（ARFN）方法，通过识别和利用图像中的情感区域，增强多模态方面级情感分类。GLFFCA(Wang et al., 2024)：一种基于全局-局部特征融合的共同注意力（GLFFCA）的多模态方面级情感分析方法。

### 4.4 在TWITTER-15和TWITTER-17数据集上的实验结果

我们在两个MABSA基准数据集上进行比较实验，以展示我们提出方法的有效性，如表3所示。我们可以看到：1) 两个基准数据集上的实验结果显示，我们提出的多模态方法优于大部分现有方法，与次优模型相比F1值在两个数据集上分别提高了0.1%和0.57%，这说明了我们提出的方法的有效性。2)与单模态方法相比，本文提出的方法显著优于所有单模态方法上，例如仅文本的方面级情感分析方法和仅图像的方面级情感分析方法。实验结果证明了仅文本的方法可以有效的利用视觉模态来提高情感分析的性能。3)与其他多模态方面级情感分析方法相比，我们提出的方法在所有评估指标上都比最近的多模态融合方法有了显着改进，这表明方面引导的语义对齐和多模态融合有助于提高情感分析性能，这确认了我们提出方法的优越性。

### 4.5 在MASAD数据集上的实验结果

为了进一步证明本文方法的鲁棒性，我们在新发布的MASAD数据集上进行了实验。表4中的实验结果显示：1) 本文的方法取得了最好的效果，在准确率ACC和F1分数上均显著超过其他多模态方面级情感分析方法。2) 在MASAD数据集的七种方面术语分类上，本文方法的结果全部优于其他多模态方面级情感分析方法。在MASAD数据集上的实验结果进一步证明了我们提出方法的有效性和鲁棒性。此外，值得注意的是，MASAD数据集是一个最近才发布的数据集，因此只有少数MABSA方法在此数据集上进行了情感分析性能评估。

### 4.6 消融实验

#### 4.6.1 模型不同组件的消融分析

为了进一步验证本文提出方法的有效性，我们在TWITTER-15和TWITTER-17两个数据集上进行了不同组件的消融实验，消融结果如表5所示。实验结果表明：1)移除方面引导的图文对比学习的模型，相比于完整的模型，在两个数据集上的两个评估指标都略微下降，这表明方面引导的图文对比学习一定程度上弥合了图文模态的差异。2)移除基于对比的跨模态语义交互模块的模型，相比于完整的模型，在两个数据集上的两个评估指标都有所下降，这表明方面引


(c) 2024 中国中文信息学会计算语言学专业委员会

| Modality | Methods | TWITTER-15 | | TWITTER-17 | |
|---|---|---|---|---|---|
| | | Acc | Macro-F1 | Acc | Macro-F1 |
| Image | Res-Aspect | 59.49 | 47.79 | 57.86 | 53.98 |
| Text | MGAN | 71.2 | 64.2 | 64.8 | 61.5 |
| | BERT | 74.3 | 70.0 | 68.9 | 66.1 |
| | BERT+BL | 74.3 | 70.0 | 68.9 | 66.1 |
| | BART | 76.0 | 67.6 | 69.5. | 67.0 |
| Text+Image | Res-MGAN | 71.7 | 63.9 | 66.4 | 63.0 |
| | Res-Bert | 75.02 | 69.21 | 69.2 | 66.48 |
| | MIMN | 71.84 | 65.69 | 65.88 | 62.99 |
| | ESAFN | 73.38 | 67.37 | 67.83 | 64.22 |
| | ViLBERT | 73.76 | 69.85 | 67.42 | 64.87 |
| | TomBERT(resnet) | 76.60 | 71.57 | 69.42 | 67.70 |
| | TomBERT(faster-rcnn) | 77.03 | 72.85 | 69.77 | 67.59 |
| | CapTrBERT | 78.01 | 73.25 | 69.77 | 68.42 |
| | saliencyBERT | 77.03 | 72.36 | 69.69 | 67.59 |
| | HIMT | 78.41 | 73.68 | 71.14 | 69.16 |
| | KEF-saliencyBERT | 78.15 | 73.54 | 71.88 | 68.94 |
| | FITE | 78.49 | 73.90 | 70.90 | 68.70 |
| | TMGM | 74.13 | 67.93 | 67.74 | 65.32 |
| | AFRN | 78.50 | 73.70 | 70.58 | 68.43 |
| | GLFFCA | 77.72 | 74.21 | 71.15 | 69.45 |
| | **Our model** | **77.92** | **74.31** | **71.22** | **70.02** |

表 3. 在TWITTER-15和TWITTER-17数据集的比较结果

| Modality | Methods | Food | | Goods | | Buildings | | Animal | | Human | | Plant | | Scenery | | Average | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Acc | F1 | Acc | F1 | Acc | F1 | Acc | F1 | Acc | F1 | Acc | F1 | Acc | F1 | Acc | F1 |
| Text | ATAE-LSTM | 93.06 | 86.13 | 95.31 | 94.85 | 94.32 | 94.64 | 93.56 | 92.85 | 91.55 | 91.34 | 93.05 | 93.4 | 91.87 | 91.03 | 93.25 | 92.03 |
| | IAN | 93.56 | 88.3 | 95.77 | 95.3 | 94.78 | 94.98 | 94.05 | 93.32 | 92.04 | 92.13 | 93.85 | 93.71 | 92.33 | 91.87 | 93.77 | 92.8 |
| | RAM | 93.64 | 89.25 | 95.59 | 95.12 | 94.57 | 94.71 | 94.29 | 93.55 | 91.75 | 91.87 | 93.5 | 93.37 | 92.15 | 91.39 | 93.64 | 92.75 |
| | TNet | 94.34 | 90.18 | 95.88 | 95.34 | 95.04 | 94.9 | 94.76 | 93.87 | 92.27 | 92.16 | 94.38 | 94.21 | 92.76 | 91.98 | 94.2 | 93.23 |
| Text+Image | MIMN | 94.72 | 91.39 | 95.93 | 95.87 | 96.26 | 95.8 | 95.03 | 94.06 | 92.54 | 92.31 | 95.04 | 94.97 | 93.17 | 92.38 | 94.67 | 93.83 |
| | TomBERT | 95.56 | 91.8 | 96.05 | 96.1 | 96.53 | 96.04 | 95.62 | 94.78 | 92.67 | 94.97 | 95.67 | 95.3 | 93.63 | 92.94 | 95.1 | 94.21 |
| | MMAP | 95.75 | 92.89 | 96.55 | 96.44 | 96.86 | 96.85 | 95.92 | 95.62 | 92.74 | 92.74 | 97.02 | 96.97 | 94.57 | 94.15 | 95.63 | 95.09 |
| | **Our model** | **96.43** | **95.81** | **97.57** | **97.06** | **97.32** | **97.29** | **99.36** | **99.56** | **96.84** | **96.75** | **97.48** | **97.59** | **95.34** | **94.37** | **97.19** | **96.92** |

表 4. 在MASAD数据集的比较结果

导的的跨模态语义交互可以为情感分析提供更精确的信息。3)移除移除基于对比的多模态语义融合模块，相比于完整的模型，在两个数据集上ACC和F1评估指标都大幅下降，这表明方面引导的跨模态语义交互有助于指导图文多模态对齐融合，提升情感分析的性能。通过以上的消融实验对比分析，验证了本文模型不同组件的有效性。

### 4.6.2 视觉信息的消融分析

为了验证本文提出方法中视觉信息的有效性，我们通过将与文本相关的图像替换为空白图像或随机选取的图像来进行实验，实验结果如表6 所示。实验结果表明，在TWITTER-15和TWITTER-17数据集上，当我们将与文本相关的图像替换为空白图像或随机选取的图像时，情感分类性能显著下降。这证明了我们提出的多模态融合方法的有效性。


(c) 2024 中国中文信息学会计算语言学专业委员会

| Method | TWITTER-15 | | TWITTER-17 | |
|---|---|---|---|---|
| | Acc | Macro-F1 | Acc | Macro-F1 |
| Full model | **77.92** | **74.31** | **71.22** | **70.02** |
| w/o Contrastive | 77.46 | 73.68 | 69.77 | 68.58 |
| w/o Interaction | 77.53 | 72.66 | 68.80 | 68.13 |
| w/o Fusion | 76.39 | 71.60 | 67.62 | 67.64 |

表 5. 在TWITTER-15和TWITTER-17数据集上模型不同组件的消融实验，w/o Contrastive是移除方面引导的图文对比学习的模型，w/o Interaction是移除基于对比的跨模态语义交互模块的模型，w/o Fusion 是移除基于对比的多模态语义融合模块的模型，Full model是本文完整的模型。

| Method | TWITTER-15 | | TWITTER-17 | |
|---|---|---|---|---|
| | Acc | Macro-F1 | Acc | Macro-F1 |
| | **77.92** | **74.31** | **71.22** | **70.02** |
| The blank image | 77.34 | 71.78 | 68.59 | 67.75 |
| The random image | 76.93 | 72.23 | 67.47 | 67.43 |

表 6. 在TWITTER-15和TWITTER-17数据集上视觉信息的消融实验

### 4.7 实例分析

| | | | |
|---|---|---|---|
| Image and text |  (a)**Happy New Year**! Watching San Jose - Anaheim replay ... believe it or not, this was the scene in **Riverside County** today. |  (b)Embattled Metro Councilman **Dan Johnson** to debate challenger **John Wit**, an independent. # Louisville |   (c) **The Seth Leibsohn Show** TONIGHT: Attorney Sheila Polk on Legalizing # Marijuana amp **Steve Hayward** ! |
| Ground truth | (New Year, Pos) (Riverside County, Neu) | (John Wit, Neu) (Dan Johnson, Neu) | (The Seth Leibsohn Show, Neu) (Steve Hayward, Pos) |
| w/o Contrastive | (Pos ✓, Neu ✓) | (Neu ✓, Neu✓) | (Neu✓, Neg✗) |
| w/o Interaction | (Pos ✓, Neg ✗) | (Neg✗, Neu✓) | (Neu✓, Neg✓) |
| w/o Fusion | (Pos , Neg ✗) | (Neu ✓, Neg✗) | (Neu✗, Neu✗) |
| Full method | (Pos ✓, Neu ✓) | (Neu✓, Neu✓) | (Neu✓, Pos✓) |

图 3. MABSA数据集上的几个案例研究。其中，Neu、Neg 和Pos 分别代表中性、负面和正面情感。符号✔ 和✗ 分别表示正确和不正确的情感极性预测。

图3展示了Twitter-15数据集中的三个示例。我们可以看到：1）本文的完整模型准确预测了这三个示例中各个方面的情感极性。2）当移除方面引导的图文对比学习时，(c)中的Steve Hayward方面情感极性预测出现错误。这一结果证实了方面引导的图文对比学习起到了模态对齐的作用，提高了MABSA的情感分析性能。3）当移除基于对比的跨模态语义交互模块时，(a)中的Riverside County方面和(b)中的John Wit方面的情感极性预测出现错误。这一结果证实了基于对比的跨模态语义交互模块很好地进行了语义的交互对齐，提高了MABSA的情感分析性能。4）移除基于对比的多模态语义融合模块导致(a)中的Riverside County方面、(b)中


(c) 2024 中国中文信息学会计算语言学专业委员会

的Dan Johnson方面以及(c)中的The Seth Leibsohn Show方面和Steve Hayward方面的情感极性预测出现错误，这证明了基于对比的多模态语义融合模块很好的融合了面向方面的多模态特征，提高了MABSA的情感分析性能。案例研究进一步验证了我们提出的方法在多模态方面级情感分析任务中的有效性。

## 5 结论

本文提出了一种基于方面引导的图文渐进融合的多模态方面级情感分析方法，通过方面引导的跨模态语义交互和多模态语义融合来缓解图像和文本之间的模态差异，实现图文多模态的对齐和融合，提高多模态特征的判别性。与现有方法相比，在三个基准任务上取得了强有竞争力的结果，三个基准数据集上的实验结果证明了本文提出的方法的有效性和优越性。深入的分析了所提出的模块的贡献，并证明了所提出方法用于多模态方面级情感分析的有效性。在未来的工作中，我们会重点探索如何深度挖掘图像中更多有效的信息，以进一步提升多模态方面级情感分析的性能。

## 参考文献

Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *CoRR*, abs/1607.06450.

Peng Chen, Zhongqian Sun, Lidong Bing, and Wei Yang. 2017. Recurrent attention network on memory for aspect sentiment analysis. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 452–461, Copenhagen, Denmark, September. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

Feifan Fan, Yansong Feng, and Dongyan Zhao. 2018. Multi-grained attention network for aspect-level sentiment classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3433–3442.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Li Jia, Tinghua Ma, Huan Rong, and Najla Al-Nabhan. 2023. Affective region recognition and fusion network for target-level multimodal sentiment classification. *IEEE Transactions on Emerging Topics in Computing*, pages 1–11.

Zaid Khan and Yun Fu. 2021. Exploiting BERT for multimodal target sentiment classification through input space translation. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pages 3034–3042.

Svetlana Kiritchenko, Xiaodan Zhu, Colin Cherry, and Saif Mohammad. 2014. NRC-Canada-2014: Detecting aspects and sentiment in customer reviews. In Preslav Nakov and Torsten Zesch, editors, *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 437–442, Dublin, Ireland, August. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880.


(c) 2024 中国中文信息学会计算语言学专业委员会

Xin Li, Lidong Bing, Wai Lam, and Bei Shi. 2018. Transformation networks for target-oriented sentiment classification. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 946–956, Melbourne, Australia, July. Association for Computational Linguistics.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23.

Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017a. Interactive attention networks for aspect-level sentiment classification. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, page 4068–4074. AAAI Press.

Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017b. Interactive attention networks for aspect-level sentiment classification. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 4068–4074.

Tetsuya Nasukawa and Jeonghee Yi. 2003. Sentiment analysis: capturing favorability using natural language processing. In *Proceedings of the 2nd International Conference on Knowledge Capture*, K-CAP '03, page 70–77, New York, NY, USA. Association for Computing Machinery.

Huy Thanh Nguyen and Minh Le Nguyen. 2018. Effective attention networks for aspect-level sentiment classification. In *2018 10th International Conference on Knowledge and Systems Engineering (KSE)*, pages 25–30.

Chi Sun, Luyao Huang, and Xipeng Qiu. 2019. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 380–385, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2016. Effective LSTMs for target-dependent sentiment classification. In Yuji Matsumoto and Rashmi Prasad, editors, *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3298–3307, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based LSTM for aspect-level sentiment classification. In Jian Su, Kevin Duh, and Xavier Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 606–615, Austin, Texas, November. Association for Computational Linguistics.

Jiawei Wang, Zhe Liu, Victor Sheng, Yuqing Song, and Chenjian Qiu. 2021. Saliencybert: Recurrent attention network for target-oriented multimodal sentiment classification. In *Pattern Recognition and Computer Vision: 4th Chinese Conference, PRCV 2021, Beijing, China, October 29 – November 1, 2021, Proceedings, Part III*, page 3–15, Berlin, Heidelberg. Springer-Verlag.

Shunjie Wang, Guoyong Cai, and Guangrui Lv. 2024. Aspect-level multimodal sentiment analysis based on co-attention fusion. *International Journal of Data Science and Analytics*, pages 1–14, 01.

Nan Xu, Wenji Mao, and Guandan Chen. 2019. Multi-interactive memory network for aspect based multimodal sentiment analysis. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 371–378.

Hao Yang, Yanyan Zhao, and Bing Qin. 2022. Face-sensitive image-to-emotional-text cross-modal translation for multimodal aspect-based sentiment analysis. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3324–3335, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Jianfei Yu and Jing Jiang. 2019. Adapting BERT for target-oriented multimodal sentiment classification. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5408–5414.


(c) 2024 中国中文信息学会计算语言学专业委员会

Jianfei Yu, Jing Jiang, and Rui Xia. 2020. Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:429–439.

Jianfei Yu, Kai Chen, and Rui Xia. 2022a. Hierarchical interactive multimodal transformer for aspect-based multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, pages 1–1.

Jianfei Yu, Jieming Wang, Rui Xia, and Junjie Li. 2022b. Targeted multimodal sentiment classification based on coarse-to-fine grained image-target matching. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 4482–4488.

Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. 2018. mixup: Beyond empirical risk minimization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

Zhe Zhang, Zhu Wang, Xiaona Li, Nannan Liu, Bin Guo, and Zhiwen Yu. 2021. Modalnet: an aspect-level sentiment classification model by exploring multimodal data with fusion discriminant attentional network. *World Wide Web*, 24(6):1957–1974.

Fei Zhao, Zhen Wu, Siyu Long, Xinyu Dai, Shujian Huang, and Jiajun Chen. 2022. Learning from adjective-noun pairs: A knowledge-enhanced framework for target-oriented multimodal sentiment classification. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6784–6794, Gyeongju, Republic of Korea, October. International Committee on Computational Linguistics.

Jie Zhou, Jiabao Zhao, Jimmy Xiangji Huang, Qinmin Vivian Hu, and Liang He. 2021. MASAD: A large-scale dataset for multimodal aspect-based sentiment analysis. *Neurocomputing*, 455:47–58.

CCL 2024

# 基于联邦知识蒸馏的跨语言社交媒体事件检测

周帅帅[1,2]，朱恩昌[1,2]，高盛祥*[1,2]，余正涛[1,2]，线岩团[1,2]，赵子霄[1,2]，陈霖[1,2]
1.昆明理工大学，信息工程与自动化学院，昆明，650500
2.昆明理工大学，云南省人工智能重点实验室，昆明，650500
20222204121@stu.kust.edu.cn,{zhuenchang.kmust,gaoshengxiang.yn}@foxmail.com,
ztyu@hotmail.com,xianyantuan@qq.com,{zhaozixiao,20232204066}@stu.kust.edu.cn

## 摘要

社交媒体事件检测是指在从各类社交媒体的内容中挖掘热点事件。在实际情况中，由于数据稀缺，社交媒体事件检测在低资源的情况下表现较差。现有的方法主要通过跨语言知识迁移等方式来缓解低资源问题，但忽略了数据隐私问题。因此，本文提出了基于联邦知识蒸馏的跨语言社交媒体事件检测框架（FedEvent），旨在将富资源客户端知识蒸馏到低资源客户端。该框架通过结合参数高效微调技术和三组对比损失，实现非英文语义空间到英文语义空间的有效映射，并采用联邦蒸馏策略，保障数据隐私的前提下实现知识的迁移。此外，我们还设计了一套四阶段生命周期机制以适应增量场景。最后，我们在真实数据集上进行实验以证明该框架的有效性。

**关键词：** 社交媒体事件检测；低资源；联邦知识蒸馏；跨语言知识迁移

# Cross-Lingual Social Event Detection Based on Federated Knowledge Distillation

**Shuaishuai Zhou[1,2], Enchang Zhu[1,2], Shengxiang Gao*[1,2], Zhengtao Yu[1,2], Yantuan Xian[1,2], Zixiao Zhao[1,2], Lin Chen[1,2]**
1.Faculty of Information Engineering and Automation,
Kunming University of Science and Technology, Kunming 650500, China
2.Yunnan Key Laboratory of Artificial Intelligence,
Kunming University of Science and Technology, Kunming 650500, China
20222204121@stu.kust.edu.cn,{zhuenchang.kmust,gaoshengxiang.yn}@foxmail.com,
ztyu@hotmail.com,xianyantuan@qq.com,{zhaozixiao,20232204066}@stu.kust.edu.cn

## Abstract

Social event detection involves identifying trending events from various social event content. In practice, due to data scarcity, social event detection performs poorly in low-resource settings. Existing methods primarily utilize cross-lingual knowledge transfer to mitigate low-resource issues; however, these approaches often overlook data privacy concerns. Consequently, this paper introduces a cross-lingual social media event detection framework, named FedEvent, which is based on federated knowledge distillation and is aimed at distilling knowledge from high-resource clients to those with low resources. The framework employs parameter-efficient fine-tuning techniques and triple contrastive losses to effectively map non-English semantic spaces to English ones, while using a federated distillation strategy to ensure data privacy. Additionally, we have developed a four-stage lifecycle mechanism to accommodate incremental scenarios. Finally, experiments on real datasets demonstrate the effectiveness of our framework.

**Keywords:** Social event detection , Low-resource , Federated knowledge distillation , Cross-lingual knowledge transfer


(c) 2024 中国中文信息学会计算语言学专业委员会

## 1 引言

社交媒体事件检测是指从社交媒体上的内容中，挖掘引发公众广泛关注和讨论的热点事件。近年来，各类社交媒体平台迅速发展，对社交媒体事件进行检测，在控制信息传播和正确引导舆论等方面具有重要意义。然而，在实际情况中，富资源语言数据量充足，模型训练效果好。但在低资源场景下，由于资源稀缺等问题，模型无法得到充足的训练数据，导致表现不达预期。因此，如何解决社交媒体事件检测存在的数据分布不均衡问题引发了广泛关注。同时，有效的社交媒体检测也为事件演化分析(Atefeh and Khreich, 2015)、舆情分析(Peng et al., 2021)、情绪分析(Gaspar et al., 2016)、企业风险管理(Nisar and Yeung, 2018)以及政治选举预测(Marozzo and Bessi, 2018)等下游任务提供了坚实的基础。

在社交媒体事件检测方面，大多数相关研究将其转化为增量聚类问题(Aggarwal and Subbian, 2012; Hu et al., 2017; Ozdikis et al., 2017; Zhang et al., 2007)、社区检测问题(Nguyen and Jung, 2015; Liu et al., 2020a; Liu et al., 2020b; Liu et al., 2020c; Yu et al., 2017)或主题建模问题(Becker et al., 2011; Cheng et al., 2014; Cordeiro, 2012; Zhou and Chen, 2014)来处理。然而，上述方法忽略了社交消息文本中丰富的语义和结构信息。为了解决以上问题，相关研究者提出了基于图神经网络（GNN）(Chung et al., 2020; Su et al., 2022)的社交媒体事件检测方法(Cui et al., 2021; Peng et al., 2019; Peng et al., 2021; Peng et al., 2023; Ren et al., 2022a; Ren et al., 2022b)，以有效捕获社交消息文本中深层语义信息和结构信息。此外，为了实现社交媒体增量事件检测，Cao et al. (2021)提出了名为KPGNN 的方法，该方法设计了一套生命周期机制，有效解决了流式场景下事件知识保留和更新难题。上述基于GNN的社交媒体事件检测方法在富资源语言（如英语）场景下取得了显著的效果，但是在低资源语言（如越南语）场景下模型性能还有很大的提升空间。因此，如何有效利用富资源语言知识来提升低资源语言的社交媒体事件检测模型性能成为了最近的研究热点。基于此，Peng et al. (2023)提出了基于跨语言模型参数知识迁移的社交媒体事件检测方法，该方法借助预训练的跨语言模块-LNMAP(Mohiuddin et al., 2020)，实现将富资源语言（英语）的模型参数知识迁移到低资源语言中。Ren et al. (2023)提出了基于跨语言知识蒸馏的社交媒体事件检测方法，其核心思想是，利用知识蒸馏将在富资源语言上训练的教师模型中的特征知识和关系知识蒸馏到低资源语言的学生模型中，该方法在低资源场景下取得了显著的效果。上述研究尝试通过跨语言知识迁移等方式来缓解低资源问题，利用富资源语言数据来帮助低资源语言模型提升性能。然而，在实际应用中，富资源数据与低资源数据通常分布于不同社交媒体平台，现有的方法采用一种数据中心化的方式来进行富资源语言到低资源语言的的迁移，显然忽略了这一特点。由于隐私和法律法规的限制，大量的高质量数据无法在不同平台之间流动，联邦学习作为一种数据去中心化的学习架构，为这一情况提供了可能的解决方法。值得一提的是，面对大量的社交媒体事件数据，联邦学习还可以降低数据传输和存储的成本，更符合显示需求。

然而，现有的基于联邦蒸馏的框架还面临诸多挑战：**i）低资源数据缺失**：由于数据稀缺等问题的影响，单独构建低资源语言模型难度较大，在低资源场景下的社交媒体事件检测方面表现达不到预期。**ii）语言迁移难度大**：由于不同语言之间的语法和表达方式不同，分布在不同客户端的数据差异较大，导致模型漂移，从而影响知识蒸馏的效果。与此同时，富资源语言与低资源语言之间的数据共享受到诸多限制也是亟待解决的问题。**iii）社会事件持续涌现**：在现实场景中，社交媒体事件具有数据量大和增长速度快的特点，对模型接收和处理新信息的能力提出更高的要求。

针对以上问题，本文提出了一种基于联邦知识蒸馏的跨语言社交媒体事件检测方法，实现了在保护数据隐私的同时，将富资源客户端知识有效地蒸馏到低资源客户端，提高低资源语言模型的性能。具体而言，针对低资源数据稀缺的问题，我们提出了在联邦学习框架下，采用知识蒸馏的策略，使得富资源客户端的数据能够有效地迁移到低资源客户端，并指导低资源客户端的模型进行训练。同时，我们训练了一个跨语言词嵌入模型，采用参数高效微调技术和三组

*高盛祥（通信作者）：gaoshengxiang.yn@foxmail.com
**基金项目**：国家自然科学基金（U23A20388, U21B2027, 62376111, 62266028, 62266027）；云南省重点研发计划（202303AP140008, 202401BC070021, 202103AA080015）；云南省科技人才与平台计划（202105AC160018）；云南省基础研究项目（202301AT070393）.
©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

对比损失，捕捉不同语言之间的语义差异，克服了语言之间的语义空间不一致的困难。此外，面对持续涌现的社交媒体事件数据量，本文在联邦框架下引入增量的生命周期，设计了一套四阶段的生命周期机制，使得模型能够不断适应传入的新数据，持续检测新的事件，从而达到数据增量的效果。实验结果表明，在低资源场景下，我们提出的框架在社交媒体事件检测任务上取得了显著的效果。通过实验分析，我们证明了FedEven框架不仅在性能上超越了传统的单机与多机环境下的基线方法，而且在确保不共享数据的前提下，实现了跨语言知识的有效迁移。

## 2 方法

本框架包括跨语言词嵌入模块、消息表征图学习模块和联邦网络消息蒸馏模块。首先，介绍联邦跨语言社交媒体事件检测的任务定义。其次，描述跨语言词嵌入模块。接着，介绍消息表征图学习模块。最后，详述联邦网络消息蒸馏模块，模型框架图如图1所示。

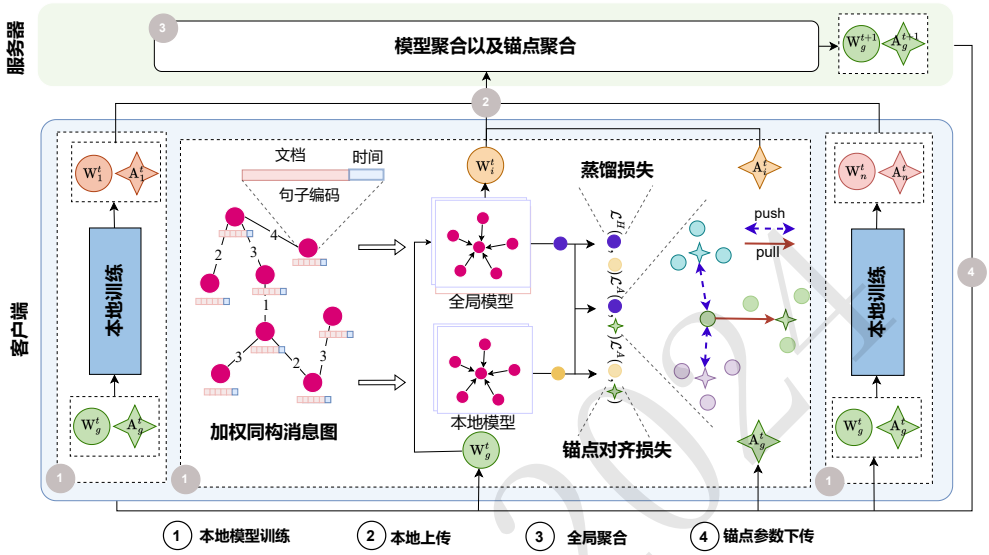



图 1. FedEvent框架图

### 2.1 任务定义

本小节将介绍本文中所使用的主要符号。我们扩展了Cao et al. (2021)中的社会消息流和社交媒体事件的定义，并给出了联邦跨语言社交事件检测的定义以及符号说明。

**定义2.1：社会消息流** $M_0, M_1, M_2, ...M_i, ...$是随时间演进的消息序列块，其中$M_i$代表在$[t_i,t_{i+1}]$期间所到达的全部社会事件$\{m_0, m_1, m_2, ...\}$，即$M_i=\{m_j|1 \le j \le |M_i|\}$，$|M_i|$是$M_i$中消息的个数。$M_0$，$\{M_i|i \neq 0\}$分别是离线场景和在线场景下的消息块。

**定义2.2：** 对于**联邦跨语言**社交媒体事件检测，低资源客户端模型可以从富资源客户端中学习有益知识从而提升模型性能。假设，H为富资源客户端集合，L为低资源客户端集合，$M_{k,i}$为第$k$个客户端上第$i$个消息块，其中每个客户端上的消息块均是单种语言。富资源客户端上第$i$个消息块中的事件集合$C\left(M_{k,i}|k \in H\right)$包含低资源客户端事件集合$C\left(M_{k,i}|k \in L\right)$，即$C\left(M_{k,i}; k \in L\right) \subseteq C\left(M_{k,i}; k \in H\right)$。低资源客户端$k$上第$t$轮次，本地模型$f_{k,l}\left(\mathbf{W}_{k,l}^{(t)}, x|x \in M_{k,i}, k \in L\right)$通过本地客户端数据集微调以及间接获取富资源客户端本地模型$f_{k,l}\left(\mathbf{W}_{k,l}^{(t)}, x|x \in M_{k,i}, k \in H\right)$的知识进行学习。

本框架的重点是在资源有限的客户端上训练一个鲁棒的模型，而非专注于训练一个能够支持多种语言的检测模型。在第$i$个消息块中，整个框架的目标是最小化所有低资源客户端的总误差，即：

$$\min_{\mathbf{w}} \Phi(\mathbf{w}) = \min_{\mathbf{w}} \sum_{\substack{k=1 \\ k \in L}}^{K} p_k \Phi_k(\mathbf{w}) \tag{1}$$

其中，$\Phi_k(\mathbf{w})$是第$k$个客户端所要优化的目标，$p_k$表示第$k$个客户端的预定义聚合权重。


(c) 2024 中国中文信息学会计算语言学专业委员会

### 2.2 跨语言词嵌入

我们利用非英文到英文的平行语料训练了一个跨语言词嵌入模型，旨在实现非英文文本到英文文本的语义空间映射。通过借助参数高效微调技术LoRA(Hu et al., 2021)的低秩自适应策略，捕捉不同语言之间的语义差异。本研究所设计的跨语言词嵌入模块（$\mathcal{M}$）是在多语言预训练模型中的注意力计算矩阵$\mathbf{W}_q$和$\mathbf{W}_v$上加一个旁支。具体而言，模型改进如下：

$$\mathbf{Q}_{\text{origin}} = \mathbf{X} \cdot \mathbf{W}_q, \mathbf{V}_{\text{origin}} = \mathbf{X} \cdot \mathbf{W}_v \tag{2}$$

$$\mathbf{Q}_{\text{lora}} = \mathbf{X} \cdot (\mathbf{W}_q + \text{ scale } \cdot \mathbf{BA}), V_{\text{lora}} = \mathbf{X} \cdot (\mathbf{W}_v + \text{scale} \cdot \mathbf{BA}) \tag{3}$$

其中，$\mathbf{Q}_{\text{origin}}, \mathbf{V}_{\text{origin}}$代表改进前$\mathbf{Q}$以及$\mathbf{V}$的计算方式，$\mathbf{Q}_{\text{lora}}, \mathbf{V}_{\text{lora}}$代表改进后$\mathbf{Q}$以及$\mathbf{V}$的计算方式。$\mathbf{B} \in \mathbb{R}^{d \times r}, \mathbf{A} \in \mathbb{R}^{r \times k}, \text{scale} = \alpha / r$，仅仅训练$\mathbf{BA}$以达到适配效果。

为了使得所引入的LoRA模块不会损失英文语义，我们引入了一个指导模型($M'$)，即被冻结的多语言预训练模型。利用对比学习来实现不同语言语义空间的映射，语义对齐损失$\mathcal{L}_{\text{total}}$为：

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{c1} + \mathcal{L}_{c2} + \mathcal{L}_{c3} \tag{4}$$

其中，$\mathcal{L}_{c1}$，$\mathcal{L}_{c2}$，和$\mathcal{L}_{c3}$ 三组对比学习损失，具体形式为：

$$\mathcal{L}_{c1} = -\sum_{i=1}^{N} \log \frac{\exp(\text{sim}(\mathcal{M}(x_i), \mathcal{M}'(y_i))/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(\mathcal{M}(x_i), \mathcal{M}'(y_j))/\tau)} \tag{5}$$

$$\mathcal{L}_{c2} = -\sum_{i=1}^{N} \log \frac{\exp(\text{sim}(\mathcal{M}(y_i), \mathcal{M}'(y_i))/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(\mathcal{M}(y_i), \mathcal{M}'(y_j))/\tau)} \tag{6}$$

$$\mathcal{L}_{c3} = -\sum_{i=1}^{N} \log \frac{\exp(\text{sim}(\mathcal{M}(x_i), \mathcal{M}(y_i))/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(\mathcal{M}(x_i), \mathcal{M}(y_j))/\tau)} \tag{7}$$

其中，$N$是一个batch中样本的数量，$\text{sim}(a, b)$表示$a$和$b$之间的相似度，$x$代表非英文文本，$y$代表相应的英文文本，$\tau$ 是温度参数，用于调节相似度的尺度。与其他自然语言处理任务一致，我们使用模型所输出的第一个TOKEN的表征（即CLS）作为整个句子的表征。该损失函数的目标是最大化正样本对之间的相似度，同时最小化与其他非对应样本之间的相似度。

### 2.3 消息表征图学习模块

#### 2.3.1 加权同构消息图构建

为了充分利用社交数据中的关键信息，我们依据用户、实体和hashtag这三种元素，构建加权同构消息图。具体而言，当两条消息共享上述三种元素中的任意一种时，就在图中为它们添加一条边。直观而言，共享的元素数量越多，两条消息之间的关系就越紧密。因此，以共现元素的个数来作为边的权重，构建加权同构消息图。在初始消息的编码方面，我们遵循了Cao et al. (2021)中结合语义和时间信息的方法。对于时间信息，我们将时间戳转换为OLE日期，进而得到二维时间特征。对于语义信息，我们采用2.2节训练的跨语言词嵌入模块对消息进行统一编码，得到语义表征。最终消息表征是消息的语义表征以及时间特征在最后一个维度上的拼接。

#### 2.3.2 GNN编码器

模型选择方面，我们选择GNN模型来作为特征提取器，主要原因有以下两点：1）GNN保留了强大的结构信息以及语义信息，并可以充分利用社交媒体上的用户、实体、hashtag元素信息。2）图神经网络较少的参数，可以更灵活的适应各个客户端，减少联邦通信时间。为了深度融合丰富的语义和关系信息，GNN编码器通过迭代地从其一跳邻居处汇总信息，学习结点表示。形式上，第$l$层消息$m_i$表示为$\boldsymbol{h}_{m_i}$，其在$l+1$层的更新可表示为：

$$\boldsymbol{h}_{m_i}^{(l+1)} \leftarrow \overset{\text{heads}}{\Big\|} \left( \boldsymbol{h}_{m_i}^{(l)} \oplus \underset{\forall m_j \in \mathcal{N}(m_i)}{\text{Aggregator}} \left( \text{Extractor} \left( \boldsymbol{h}_{m_j}^{(l)} \right) \right) \right) \tag{8}$$

CCL 2024

其中，$\mathcal{N}(m_i)$代表消息$m_i$的一跳邻居。$\oplus$代表特征聚合方式，$\overset{\text{heads}}{\Big\|}$代表多个头的拼接。聚合器（Aggregator）以及抽取器（Extractor）在不同的GNN模型中是不同的。不同的客户端可以根据本地数据设计不同的GNN，进行本地模型的私有化定制。

### 2.4 联邦网络消息蒸馏模块

在整体架构上，我们遵循了传统的基于知识蒸馏的服务器-客户端设计。服务器端，我们设计了一个全局模型$\mathbf{W}_g$和全局锚点$\mathbf{A}_g$来进行富资源客户端以及低资源客户端之间的通信。每个客户端包含一个本地模型$\mathbf{W}_{k,l}$用于适应本地数据、一个全局模型$\mathbf{W}_{k,g}$用于特征蒸馏、一个全局锚点$\mathbf{A}_{k,g}$用于对齐各个客户端特征空间。我们利用全局模型作为过程监督，全局锚点作为结果监督来实现知识的有效迁移。其中，本地模型根据本地数据进行更新以适应本地数据特点。全局模型在所有客户端之间共享，并进行协作学习。我们利用生命周期机制来适应在线以及离线场景。接下来，我们将详细介绍联邦聚合方式以及生命周期机制。

#### 2.4.1 联邦聚合

在生命周期不同阶段，联邦聚合方式略有不同，但也有一定的共性。我们假设当前块为$M_i$，以此块为例，介绍全局锚点聚合以及全局模型聚合。

**本地锚点计算：**本地锚点由当前客户端上消息块$M_i$中每个簇内所有样本表征的均值来表示。因此，本地锚点是当前客户端上源自同一事件消息簇特性的综合体现。$t$轮次时，客户端$k$中事件簇c的局部锚点表征$a_{k,c}^{(t)}$计算如下：

$$a_{k,c}^{(t)} = \frac{1}{|\mathcal{B}_{k,c}|} \sum_{(\mathbf{x},c)\in\mathcal{B}_{k,c}} f_{k,l}\left(\mathbf{W}_{k,l}^{(t)}, \mathbf{x}\right) \tag{9}$$

其中，$|\mathcal{B}_{k,c}|$表示客户端$k$上事件簇c的样本数量。$f_{k,l}$代表本地模型的输出。

**全局锚点聚合：**中央服务器汇总来自参与聚合的客户端上本地锚点表征，聚合方式根据各个客户端消息块中消息个数进行加权平均，即：

$$a_c^{(t)} := \frac{1}{\sum_{k\in S}|\mathcal{B}_{k,c}|} \sum_{k\in S} |\mathcal{B}_{k,c}|\, \mathbf{a}_{k,c}^{(t)} \tag{10}$$

其中，S为满足聚合条件的客户端集合，在生命周期机制中不同阶段，集合S有所不同，我们将在2.4.2节进行说明。通过公式(10)得到全局锚点$A_g^{(t+1)} = \{a_c^{(t)}\}_{c=1}^{C}$将会被广播到所共享的各个客户端。此时，所获得的全局锚点能够综合反映同一事件簇的特征，可以用来指导每个客户端的模型训练。

**全局模型聚合：**中央服务器接收并汇总来自所有客户端上的全局模型，通过聚合处理生成下一轮更新的全局模型$\mathbf{W}_g^{(t+1)}$，即：

$$\mathbf{W}_g^{(t+1)} := \frac{1}{\sum_{k\in S}|\mathcal{B}_k|} \sum_{k\in S} |\mathcal{B}_k|\, \mathbf{W}_{k,g}^{(t)} \tag{11}$$

其中，$|\mathcal{B}_k|$表示客户端$k$中样本个数，$\mathbf{W}_{k,g}^{(t)}$为第$k$个客户端中第$t$轮时全局模型的参数。

在每一轮通信中，当本地训练完后，客户端$k$将$\{\mathbf{A}_k^{(t)}, \mathbf{W}_{k,g}^{(t)}\}$上传至中央服务器，通过公式(10),(11)进行聚合，然后将聚合结果$\{\mathbf{A}_g^{(t+1)}, \mathbf{W}_g^{(t+1)}\}$下拉至各个客户端，进行新一轮的本地训练。

#### 2.4.2 生命周期机制

生命周期图如图2所示。FedEvent的生命周期包括富资源客户端预训练(Stage Ⅰ)、低资源客户端预训练(Stage Ⅱ)、各个客户端检测(Stage Ⅲ)以及维护(Stage Ⅳ)四个阶段。在第一阶段，我们利用富资源客户端消息块上$M_0$构建一个初始的加权同构消息图并利用特定的算法训练初始模型。在第二阶段，我们利用阶段一所训练的模型以中央服务器作为知识传输的媒介来辅助低资源客户端训练初始模型。在第三阶段，为每个到来的块构建一个新的图，然后使用已训练好的模型直接检测每个到来的消息块。在第四阶段，持续用最新的消息块训练模型，并利用联邦


(c) 2024 中国中文信息学会计算语言学专业委员会

共蒸馏机制来实现各个客户端在线学习，使模型能够学习新知识。其中，阶段一以及阶段二用于离线学习，阶段四用于在线学习。阶段三及阶段四交替进行，通过这种方式，模型不断适应传入的数据，可以检测新事件并更新模型的知识。通过生命周期机制，FedEvent能够更好地适应流媒体特性。

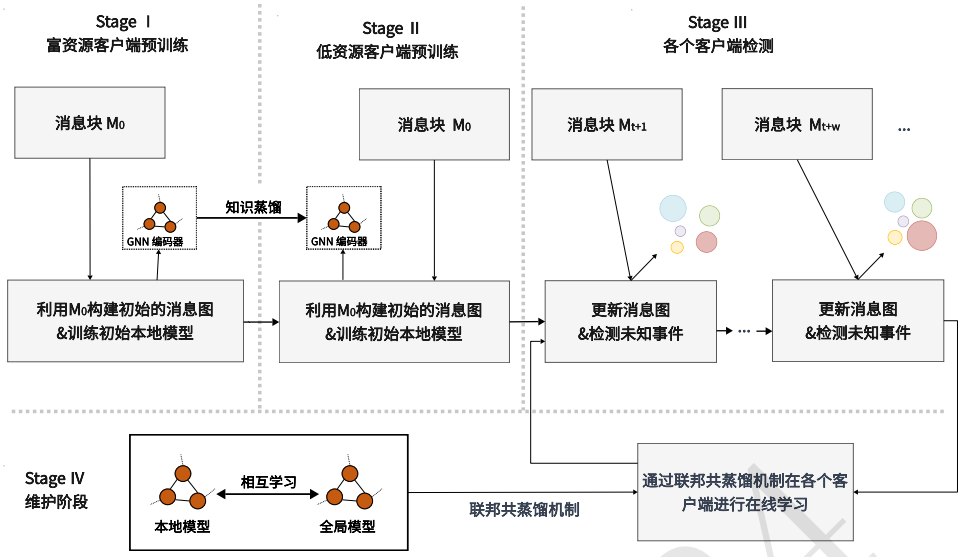



图 2. FedEvent生命周期机制

**Stage I：富资源客户端预训练。**该阶段的目标是在富资源客户端利用初始消息块$M_0$训练一个鲁棒的本地模型$\mathbf{W}_{k,l}$。随着新事件的不断涌现，事件的总数很难提前知道。考虑到交叉熵损失只适用于事件总数是已知和固定的情况，难以处理流式数据。因此，我们转向利用成对学习方法(Ren et al., 2022a)，保证簇间紧密性和簇内分离性。由于这种特性，使得在后续第二、三阶段，可以将富资源客户端清晰的簇间界限信息蒸馏至低资源客户端。该阶段任务特定损失以及总损失分别为$\mathcal{L}^P$，$\mathcal{L}_1$:

$$\mathcal{L}_1 = \mathcal{L}^P = \sum_{\substack{(m_i,m_{i+})\in\{Pos\}\\(m_j,m_{j-})\in\{Neg\}}} \max\{\mathcal{D}(\boldsymbol{h}_{m_i},\boldsymbol{h}_{m_{i+}}) - \mathcal{D}(\boldsymbol{h}_{m_j},\boldsymbol{h}_{m_{j-}}) + a, 0\} \tag{12}$$

其中，$\mathcal{L}_1$是第一阶段损失，$\{Pos\}$是来自相同事件簇的样本对，$\{Neg\}$是来自不同事件簇的样本对。$a$是边界距离超参数，$D(,)$计算欧氏距离，$\boldsymbol{h}$是由GNN编码器所学习到的消息表征。

**Stage Ⅱ：低资源客户端预训练。**该阶段目标是利用阶段一所训练出的本地模型来协助低资源客户端训练初始模型。我们提出了以全局模型为导向的的过程监督及以全局锚点为导向的结果监督来实现知识的迁移。具体来说，对于富资源客户端，冻结本地模型，并以本地模型作为教师模型，全局模型作为学生模型。学生模型学习富资源客户端数据特征的同时，也将学习教师模型的隐藏层输出（Local-To-Global策略）。在低资源客户端，冻结全局模型，以全局模型作为教师模型，本地模型作为学生模型，学生模型从教师模型中学习通用知识的同时，利用本地数据学习适合本地数据特征的知识（Global-To-Local策略）。我们将以上两种策略作为以过程为导向的监督。并引入以全局锚点为导向的结果监督来统一各个客户端的簇心。从结果导向和过程导向两个方面进行知识迁移。以下将介绍这两种监督方式及该阶段总损失。

在以过程为导向的监督学习中，学生模型可以根据本地数据特点，灵活地学习教师模型中特定或若干隐藏层的输出。客户端$k$上，各个客户端学生模型蒸馏损失$\mathcal{L}_k^{KD}$可表示为:

$$\mathcal{L}_k^{KD} = \frac{1}{I}\sum_{i=1}^{I} \mathrm{MSE}(\mathbf{H}_k^{t,i}; \mathbf{H}_k^{s,i}) \tag{13}$$

其中，MSE表示均方误差，I代表学生模型要向教师模型学习的隐藏层个数，$\mathbf{H}_k^{t,i}, \mathbf{H}_k^{s,i}$ 分别代表教师模型、学生模型第$i$个隐藏层的输出。


(c) 2024 中国中文信息学会计算语言学专业委员会

在本地客户端训练时，我们期待每个特征可以与相应的全局锚点更接近，同时远离非相应的全局锚点，我们利用对比学习来协调所有客户端的本地中心，以便在客户之间交换信息。全局锚点以及客户端上模型可分别表示为$\mathbf{A}_k^{(t)} = \{a_c^{(t)}\}_{c=1}^{C}, \mathbf{W}_k^{(t)}$，对于第$i$块的数据采样$x$，锚点对齐损失$\mathcal{L}_k^A$为：

$$\mathcal{L}_k^A = \mathcal{L}_{CE}(s, c) \tag{14}$$

$$s_n = \frac{\exp\left(\left\langle \mathbf{a}_n^{(t)}, f\left(W_k^{(t)}, x\right)/\alpha \right\rangle\right)}{\sum_{i=1}^{C} \exp\left(\left\langle \mathbf{a}_i^{(t)}, f\left(W_k^{(t)}, x\right)/\alpha \right\rangle\right)} \tag{15}$$

其中$\mathcal{L}_{CE}$为交叉熵损失，$s = [s_1, s_2, \ldots, s_c]$是一个相似度向量，$s_n$ 可表示当前特征与各个锚点的相似度，$\alpha$决定浓度的水平，$\langle,\rangle$是内积。最小化交叉熵损失既最大限度地提高特征与其对应的锚点$a_c^{(t)}$之间的相似性，又最小化该特征与$\{a_n^{(t)}\}_{n\neq c}$中每个非对应锚点之间的相似性。

阶段二中参与聚合的客户端集合S=H（公式(10),(11)），即仅仅富资源客户端参与聚合。客户端$k$上，该阶段仅仅训练各个客户端上学生模型，该阶段总损失$\mathcal{L}_2$可表示为：

$$\mathcal{L}_2 = \mathcal{L}_k^P + \mathcal{L}_k^{KD} + \mathcal{L}_k^A \tag{16}$$

**Stage Ⅲ：各个客户端检测。**在检测阶段，我们根据所训练出来的模型对$[M_i, M_{i+w}]$窗口内的消息块进行聚类。基于距离的聚类算法（如K-Means）或基于密度的算法（如HDBSCAN(Campello et al., 2013)）都可以对消息进行聚类。其中，HDBSCAN不需要指定类别的总数，非常适合增量的检测需求。因此，我们利用HDBSCAN算法进行消息检测。

**Stage Ⅳ：各个客户端联合训练（维护阶段）。**该阶段，受Wu et al. (2022)工作的启发，我们设计了基于任务特定损失引导的共蒸馏机制来适应在线的场景。在此过程中，各个客户端的本地模型以及全局模型均参与模型训练，相互学习并适应在线场景下的数据分布和模式变化。由于本地模型比全局模型更加适应本地数据，本地模型编码的有用知识可以帮助指导全局模型。此外，本地模型只能从本地数据中学习，而全局模型融合了所有客户端的数据特性。因此，本地模型还可以从全局模型蒸馏的知识中受益。所以，本地模型和全局模型之间可以相互学习。基于此，该阶段我们设计了基于任务特定损失引导的共蒸馏机制，使本地模型及全局模型相互学习对方有用的知识。针对于全局模型和本地模型，蒸馏损失可表示为：

$$\mathcal{L}_{k,l}^H = \mathcal{L}_{k,g}^H = \frac{\frac{1}{I}\sum_{i=1}^{I} \mathrm{MSE}(\mathbf{H}_k^{t,i}; \mathbf{H}_k^{s,i})}{\mathcal{L}_{k,l}^P + \mathcal{L}_{k,g}^P} \tag{17}$$

该阶段中参与聚合的客户端集合$S = \{L \cup H\}$(公式(10),(11))，即所有客户端均参与聚合。该阶段总损失包含任务特定损失、基于任务特定损失的共蒸馏损失以及锚点对齐损失。该阶段本地模型、全局模型的损失分别可表示为$\mathcal{L}_{l,3}$、$\mathcal{L}_{g,3}$。

$$\mathcal{L}_{l,3} = \mathcal{L}_{k,l}^P + \mathcal{L}_{k,l}^H + \mathcal{L}_{k,l}^A \tag{18}$$

$$\mathcal{L}_{g,3} = \mathcal{L}_{k,g}^P + \mathcal{L}_{k,g}^H + \mathcal{L}_{k,g}^A \tag{19}$$

其中，$\mathcal{L}_{l,3}$、$\mathcal{L}_{g,3}$分别代表该阶段本地模型及全局模型的总训练损失。$\mathcal{L}_{k,l}^P$、$\mathcal{L}_{k,g}^P$分别代表该阶段本地模型和全局模型的任务特定损失（公式(12)）。$\mathcal{L}_{k,l}^H$、$\mathcal{L}_{k,g}^H$分别代表该阶段本地模型和全局模型的共蒸馏损失（公式(17)）。$\mathcal{L}_{k,l}^A$、$\mathcal{L}_{k,g}^A$分别代表该阶段本地模型和全局模型的锚点对齐损失（公式(14)）。

## 3 实验

### 3.1 数据集及评价指标

(1) 实验数据集：

为了评估FedEvent框架的性能，本研究在一个大型公开的英文数据集Events2012(McMinn et al., 2013)的基础上，根据其主题描述补充了中文和越南语事件样本，构建了低资源客户端上私有数据集。这项工作旨在创建一个多客户端、跨语言的实验环境。首先，我们依据twitter高


(c) 2024 中国中文信息学会计算语言学专业委员会

质量筛选器指南[1]，通过关键词、时间范围和语言限制进行初步的数据筛选。例如，使用查询语句“(莫言AND 奖) since:2012-10-12 until:2012-10-13 lang:zh” 可以获得2012年10月12日至13日期间，包含关键词“莫言”和“奖”的中文推文。接下来，我们通过人工审核和数据清洗过程来确保数据的质量。对于文本实体的抽取，我们分别采用了spacy[2](对于中文和英文)和Underthesea[3]（对于越南语）。该数据集各个消息块统计信息如表1所示。

| 消息块 | $M_0$ | $M_1$ | $M_2$ | $M_3$ | $M_4$ | $M_5$ | $M_6$ | $M_7$ | $M_8$ | $M_9$ | $M_{10}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 中文消息个数 | 3765 | 1151 | 348 | 468 | 316 | 405 | 300 | 1022 | 108 | 132 | 238 |
| 越南语消息个数 | 2716 | 1270 | 207 | 239 | 244 | 249 | 145 | 743 | 181 | 146 | 113 |
| 消息块 | $M_{11}$ | $M_{12}$ | $M_{13}$ | $M_{14}$ | $M_{15}$ | $M_{16}$ | $M_{17}$ | $M_{18}$ | $M_{19}$ | $M_{20}$ | $M_{21}$ |
| 中文消息个数 | 234 | 297 | 204 | 714 | 248 | 175 | 757 | 346 | 290 | 101 | 267 |
| 越南语消息个数 | 140 | 426 | 264 | 387 | 325 | 113 | 350 | 246 | 155 | 80 | 292 |

表 1. 数据集统计情况

(2)评价指标:

本研究采用标准化互信息（NMI）和调整互信息（AMI）作为评价指标，以确保与以往社交媒体事件检测工作的可比性。

### 3.2 实验参数设置

在本研究中，跨语言模块训练基于Hungging Face平台提供的平行语料库[4]，其中，LoRA的秩$r$设置为4。为保证实验的公正性，除mBERT外的所有基线方法均使用同一跨语言模块($\mathcal{M}$)进行编码。全局模型采用双层图注意力网络（GAT），注意力头设置为4，隐藏层和输出层维度均设为32。第一阶段，各个客户端本地迭代次数设置为2，轮次设置为10。第二阶段，轮次减至5。第三阶段中窗口大小$w$设为3。第四阶段，轮次设置为10。各个阶段中模型学习率设置为0.001，批次大小设置为1000，成对损失的系数$\alpha$取10，优化器选用Adam(Kingma and Ba, 2014)。所有实验重复5次，来记录结果的均值。聚类算法HDBSCAN中核心样本的密度阈值（min_samples）设置为3，最小聚类大小(min_cluster_size)设置为10。

### 3.3 基线方法

本文将所提出的FedEvent方法与下述社交事件检测方法进行离线和在线场景下的比较。

• mBERT(Pires et al., 2019): 利用mBERT(Pires et al., 2019)得到的768维句子嵌入作为消息的表示，以捕捉深层次语义信息。

• KPGNN(Cao et al., 2021): 通过图注意力网络（GAT）结合三元损失函数获取语义及结构上的消息表示。

• QSGNN(Ren et al., 2022a):通过最佳的已知样本和可靠的知识转移，将知识从已知扩展到未知。

• FedKP: 我们将FedAvg(McMahan et al., 2017)和KPGNN(Cao et al., 2021)进行结合，让每个客户端都拥有一个KPGNN主干模型。

• FedQS: 我们将FedAvg(McMahan et al., 2017)中各个客户端主干模型设为QSGNN(Ren et al., 2022a)。

• CLKD(Ren et al., 2023): 利用中心化的训练方式将富资源语言上训练的教师模型中的特征知识以及关系知识蒸馏到低资源语言的学生模型中。

### 3.4 主实验

为了评估本文提出的模型的性能，我们在每个客户端上进行了离线评估以及在线评估。

[1]https://developer.twitter.com/zh-cn/docs/tutorials/building-high-quality-filters

[2]https://spacy.io/api/annotation#section-named-entities

[3]https://underthesea.readthedocs.io/en/latest/readme.html

[4]越英平行语料:https://huggingface.co/datasets/larryvrh/WikiMatrix-v1-En_Zh-filtered，中英平行语料:https://huggingface.co/datasets/harouzie/vi_en-translation


(c) 2024 中国中文信息学会计算语言学专业委员会

我们使用该客户端数据集中第一周的消息构建初始消息块$M_0$，并将随后21天的数据划分成消息块$M_1, M_2, \ldots, M_{21}$。我们在这两种场景下做了对比实验以及消融实验。实验结果记录在表2,3,4,5中。

| Metrics | $mBERT_c$ | | $KPGNN_c$ | | $QSGNN_c$ | | $FedKP_c$ | | $FedQS_c$ | | $\mathcal{M}_a$ | | $w/o\ NC_a$ | | $w/o\ NA_a$ | | $FedEvent$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | zh | vi | zh | vi | zh | vi | zh | vi | zh | vi | zh | vi | zh | vi | zh | vi | zh | vi |
| NMI | .23 | .01 | .59 | .42 | .62 | .44 | .63 | .47 | .64 | .53 | .41 | .13 | .66 | .45 | .68 | .60 | **.70** | **.63** |
| AMI | .14 | .01 | .47 | .31 | .51 | .35 | .52 | .37 | .53 | .36 | .30 | .12 | .53 | .34 | .58 | .45 | **.62** | **.45** |

表 2. 低资源客户端离线评估结果。其中，zh代表中文客户端，vi代表越南语客户端，带脚标$c$的方法用于对比实验，带脚标$a$的方法用于消融实验。$w/o\ NC_a$、$w/o\ NA_a$分别代表本方法中去除跨语言词嵌入模块、锚点对齐损失。

**离线评估**：对于每个客户端，我们将该客户端上的初始消息块$M_0$按照70%、20%和10%划分为训练集、测试集以及验证集。从表2中带有脚标$c$的实验结果可以看出，在离线场景下，多机环境（FedEvent,FedKP等）优于单机环境下的事件检测方法，证明了多机环境可以提升单机环境下模型的性能。究其原因，是因为联邦框架可以整合来自各个客户端上的知识。具体而言，在中文客户端上，多机环境相比于单机环境下的方法在NMI以及AMI上分别表现出了1.6%-204.3%，2.0%-342.9%的显著提升，在越南语客户端上，分别提升了2.3%-6200%，0%-4400%。引入联邦后，FedKP、FedQS方法虽然效果也有提升，但并不明显，是因为各个客户端上的特征空间不一致以及不同语言之间的差异。而FedEvent利用全局锚点来统一各个客户端上的特征空间，利用本地模型来适应本地私有数据的特性，并利用全局模型来传递全局的知识。所以，FedEvent在处理分布式社交事件检测任务时，更适合来实现富资源客户端知识迁移至低资源客户端。具体而言，与FedKP、FedQS相比，FedEvent在NMI和AMI指标上分别提升了11.1%-34.0%，17.0%-28.6%，展现了我们框架的有效性。

| Blocks | $M_1$ | | | | $M_2$ | | | | $M_3$ | | | | $M_4$ | | | | $M_5$ | | | | $M_6$ | | | | $M_7$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Metrics | NMI | | AMI | | NMI | | AMI | | NMI | | AMI | | NMI | | AMI | | NMI | | AMI | | NMI | | AMI | | NMI | | AMI | |
| | zh | vi | zh | vi | zh | vi | zh | vi | zh | vi | zh | vi | zh | vi | zh | vi | zh | vi | zh | vi | zh | vi | zh | vi | zh | vi | zh | vi |
| **mBERT$_c$** | .15 | .01 | .10 | .01 | .22 | .56 | .19 | .46 | .10 | .10 | .08 | .01 | .09 | .02 | .01 | .06 | .07 | .03 | .05 | .01 | .40 | .42 | .34 | .33 | .02 | .02 | .01 | .01 |
| **KPGNN$_c$** | .54 | .49 | .49 | .47 | .75 | .79 | .71 | .74 | .55 | .71 | .53 | .65 | .64 | .62 | .58 | .56 | .69 | .65 | .74 | .56 | .78 | .69 | .75 | .60 | .46 | .46 | .43 | .42 |
| **QSGNN$_c$** | .57 | .51 | .53 | .47 | .73 | .74 | .68 | .69 | .54 | .72 | .51 | .69 | .65 | .59 | .59 | .50 | .68 | .66 | .64 | .58 | .70 | .67 | .64 | .65 | .47 | .40 | .45 | .36 |
| **FedKP$_c$** | .58 | .55 | .54 | .53 | .74 | .77 | .69 | .73 | .58 | .80 | .55 | .78 | .65 | .65 | .58 | .59 | .74 | **.69** | .71 | **.61** | .84 | .73 | .75 | .72 | .53 | .55 | .52 | .51 |
| **FedQS$_c$** | .60 | .56 | .53 | .54 | .70 | .76 | .61 | .71 | .59 | .79 | .56 | .71 | .57 | .51 | .48 | .41 | .71 | .53 | .68 | .42 | .79 | .72 | .77 | .66 | .54 | .39 | .51 | .32 |
| $\mathcal{M}_a$ | .39 | .14 | .33 | .13 | .57 | .57 | .49 | .47 | .39 | .21 | .31 | .16 | .50 | .47 | .43 | .42 | .40 | .34 | .31 | .28 | .62 | .56 | .58 | .47 | .01 | .15 | .01 | .12 |
| **w/o NC$_a$** | .51 | .45 | .49 | .40 | .70 | .68 | .66 | .63 | .63 | .78 | .60 | .73 | .63 | .67 | .60 | .61 | .69 | .63 | .66 | .63 | .82 | .71 | .80 | .64 | .50 | .48 | .47 | .42 |
| **w/o NA$_a$** | .56 | .55 | .50 | .54 | .79 | .75 | .75 | .70 | .66 | .82 | .61 | .78 | .67 | .62 | .61 | .55 | **.75** | .67 | **.72** | .59 | .89 | .77 | .86 | .71 | .53 | .55 | .50 | .51 |
| **FedEvent** | **.61** | **.62** | **.55** | **.57** | **.81** | **.82** | **.76** | **.79** | **.70** | **.87** | **.66** | **.84** | **.70** | **.65** | **.65** | **.58** | .74 | **.69** | .69 | .60 | **.91** | **.80** | **.89** | **.73** | **.58** | **.58** | **.55** | **.54** |

表 3. 低资源客户端在线评估结果（$M_1 - M_7$）

**在线评估**：在线场景中，实验结果如表3,4,5所示，我们的方法在在线场景下大多都超过了单机环境和多机环境下的所有方法。在线评估中，我们观察到不同块之间性能差异明显，可能是因为不同消息块中消息数量有着较大差异以及消息块中数据的长尾分布。

### 3.5 对照实验

为了探究我们提出的去中心化蒸馏方法FedEvent与目前最先进的中心化蒸馏方法（CLKD）之间的性能差异，我们设计了一组对照实验。实验结果如表6所示。表格中的数据代表低资源客户端（中越）模型在消息块$M_i$上评估结果（NMI、AMI）的加权平均。实验结果表明，本文所提出的方法与最先进的中心化方法（CLKD）在性能上非常接近，总体上仅下降了0.02，这进一步证明了我们方法的显著优势。主要涉及两个方面的原因，一方面，FedEvent通过锚点对齐损失来统一各个客户端上的簇心，使得全局模型拥有全局一致的知识。在此基础上，本地模型会结合全局知识以及本地数据进行微调，从而使其更加符合本地数

| Blocks | $M_8$ | | | | $M_9$ | | | | $M_{10}$ | | | | $M_{11}$ | | | | $M_{12}$ | | | | $M_{13}$ | | | | $M_{14}$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Metrics | NMI | | AMI | | NMI | | AMI | | NMI | | AMI | | NMI | | AMI | | NMI | | AMI | | NMI | | AMI | | NMI | | AMI | |
| | zh | vi | zh | vi | zh | vi | zh | vi | zh | vi | zh | vi | zh | vi | zh | vi | zh | vi | zh | vi | zh | vi | zh | vi | zh | vi | zh | vi |
| **mBERT$_c$** | .28 | .18 | .23 | .13 | .10 | .09 | .03 | .03 | .33 | .18 | .27 | .13 | .29 | .47 | .23 | .41 | .33 | .02 | .29 | .01 | .42 | .44 | .37 | .39 | .27 | .38 | .20 | .31 |
| **KPGNN$_c$** | .72 | .75 | .65 | .66 | .74 | .71 | .69 | .64 | .80 | .82 | .86 | .79 | .74 | .52 | .70 | .50 | .75 | .61 | .70 | .53 | .75 | .75 | .74 | .72 | .41 | .58 | .37 | .52 |
| **QSGNN$_c$** | .74 | .71 | .68 | .64 | .79 | .71 | .76 | .63 | .81 | .79 | .83 | .76 | .71 | .53 | .67 | .49 | .71 | .44 | .65 | .37 | .66 | .55 | .63 | .49 | .42 | .62 | .38 | .56 |
| **FedKP$_c$** | .77 | .78 | .72 | .71 | .79 | .78 | .74 | .72 | .83 | .83 | .81 | .80 | .71 | .54 | .66 | .45 | **.81** | .68 | **.77** | .61 | .79 | .80 | .77 | .78 | .44 | .64 | .36 | .57 |
| **FedQS$_c$** | .78 | .75 | .71 | .70 | .79 | .76 | .78 | .70 | .84 | .80 | .80 | .74 | .72 | .57 | .68 | .48 | .73 | .66 | .67 | .60 | .70 | .64 | .68 | .60 | .40 | .61 | .31 | .55 |
| $\mathcal{M}_a$ | .54 | .41 | .49 | .34 | .42 | .51 | .36 | .41 | .52 | .51 | .49 | .43 | .46 | .45 | .39 | .38 | .45 | .33 | .39 | .26 | .62 | .36 | .58 | .29 | .33 | .42 | .26 | .35 |
| **w/o NC$_a$** | .65 | .75 | .60 | .66 | .83 | .77 | .80 | .72 | .79 | .80 | .77 | .78 | .70 | .59 | .65 | .51 | .76 | .62 | .72 | .52 | .77 | .78 | .75 | .75 | .43 | .57 | .37 | .53 |
| **w/o NA$_a$** | .79 | .78 | .72 | .71 | **.85** | .80 | .82 | .75 | .84 | .82 | .83 | .79 | .74 | .55 | .70 | .46 | .77 | .66 | .72 | .60 | .89 | .78 | .86 | .75 | .43 | .62 | .36 | .57 |
| **FedEvent** | **.85** | **.84** | **.79** | **.78** | **.85** | **.85** | **.82** | **.81** | **.87** | **.93** | **.86** | **.92** | **.79** | **.68** | **.72** | **.64** | .78 | **.72** | .73 | **.65** | **.91** | **.87** | **.90** | **.82** | **.47** | **.65** | **.41** | **.58** |

表 4. 低资源客户端在线评估结果（$M_8 - M_{14}$）

| Blocks | $M_{15}$ | | | | $M_{16}$ | | | | $M_{17}$ | | | | $M_{18}$ | | | | $M_{19}$ | | | | $M_{20}$ | | | | $M_{21}$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Metrics | NMI | | AMI | | NMI | | AMI | | NMI | | AMI | | NMI | | AMI | | NMI | | AMI | | NMI | | AMI | | NMI | | AMI | |
| | zh | vi | zh | vi | zh | vi | zh | vi | zh | vi | zh | vi | zh | vi | zh | vi | zh | vi | zh | vi | zh | vi | zh | vi | zh | vi | zh | vi |
| **mBERT$_c$** | .34 | .32 | .29 | .28 | .32 | .29 | .28 | .24 | .25 | .03 | .19 | .01 | .36 | .21 | .33 | .16 | .04 | .06 | .02 | .03 | .42 | .32 | .34 | .26 | .03 | .39 | .01 | .33 |
| **KPGNN$_c$** | .53 | .60 | .47 | .54 | .61 | .74 | .58 | .71 | .53 | .73 | .51 | .70 | .51 | .37 | .44 | .32 | .77 | .75 | .75 | .79 | .73 | .72 | .69 | .66 | .48 | .56 | 42 | .50 |
| **QSGNN$_c$** | .54 | .61 | .51 | .56 | .62 | .74 | .61 | .69 | .44 | .71 | .42 | .68 | .45 | .37 | .42 | .33 | .73 | .61 | .71 | .53 | .73 | .53 | .69 | .47 | .43 | .47 | .40 | .40 |
| **FedKP$_c$** | .59 | .65 | .56 | .59 | **.89** | **.85** | **.88** | **.82** | .52 | .73 | .47 | .70 | .66 | .49 | .62 | .44 | .79 | .68 | .77 | .63 | .78 | .78 | .75 | .75 | .52 | .63 | .47 | .57 |
| **FedQS$_c$** | .55 | .53 | .46 | .44 | .80 | .81 | .76 | .79 | .48 | .58 | .46 | .52 | .62 | .34 | .57 | .29 | .75 | .69 | .70 | .63 | .67 | .57 | .58 | .50 | .52 | .54 | .50 | .50 |
| $\mathcal{M}_a$ | .47 | .41 | .39 | .32 | .59 | .43 | .56 | .35 | .36 | .50 | .31 | .44 | .32 | .28 | .28 | .23 | .51 | .29 | .46 | 23 | .51 | .29 | .42 | .31 | .03 | .29 | .01 | .24 |
| **w/o NC$_a$** | .55 | .61 | .55 | .55 | .65 | .79 | .62 | .76 | .56 | .63 | .52 | .60 | .55 | .47 | .48 | .42 | .74 | .73 | .72 | .68 | .74 | .70 | .68 | .67 | .57 | .57 | .50 | .52 |
| **w/o NA$_a$** | .61 | .64 | .46 | .58 | **.89** | .77 | .87 | .73 | .56 | .75 | .53 | .70 | .70 | .58 | .67 | .49 | **.82** | .75 | **.81** | .70 | .78 | .81 | .72 | .78 | .59 | .59 | .51 | .52 |
| **FedEvent** | **.70** | **.67** | **.63** | **.60** | .83 | .80 | .80 | .76 | **.63** | **.83** | **.58** | **.78** | **.73** | **.60** | **.69** | **.51** | .75 | **.83** | .73 | **.80** | **.80** | **.88** | **.77** | **.83** | **.62** | **.66** | **.55** | **.63** |

表 5. 低资源客户端在线评估结果（$M_{15} - M_{21}$）

据的特性。另一方面，本方法采用了自适应蒸馏策略，来选择性的获取对方有益的知识，从而避免对无益的知识进行传输。

| **Blcoks** | $M_0$ | $M_1$ | $M_2$ | $M_3$ | $M_4$ | $M_5$ | $M_6$ | $M_7$ | $M_8$ | $M_9$ | $M_{10}$ | $M_{11}$ | $M_{12}$ | $M_{13}$ | $M_{14}$ | $M_{15}$ | $M_{16}$ | $M_{17}$ | $M_{18}$ | $M_{19}$ | $M_{20}$ | $M_{21}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **CLKD** | .57 | .57 | .79 | **.78** | **.65** | **.69** | .82 | **.58** | **.84** | **.85** | .86 | **.72** | .72 | .86 | **.55** | **.67** | **.82** | **.72** | .62 | **.78** | .81 | .61 |
| **FedEvent** | **.60** | **.59** | **.80** | .77 | **.65** | .68 | **.83** | .56 | .81 | .83 | **.90** | .71 | **.72** | **.87** | .53 | .65 | .80 | .71 | **.63** | **.78** | **.82** | **.62** |

表 6. 去中心化与中心化蒸馏方法对照实验结果（$M_0 - M_{21}$）

### 3.6 消融实验

为了验证本文方法中各个模块的有效性，本文设置了以下两组消融实验：跨语言词嵌入模块和锚点对齐损失对模型性能的影响。通过对比表2,3,4,5中带有脚标$a$的实验结果，我们观察到，无论是在离线还是在线场景下，跨语言词嵌入模块和锚点对齐损失都显著提升了模型性能。跨语言词嵌入模块$\mathcal{M}$与mBERT相比，$\mathcal{M}$明显缓解了mBERT在低资源语言文本表征上的不足，具体而言，在中文客户端上，跨语言词嵌入模块$\mathcal{M}$相比于mBERT在NMI和AMI指标上平均提升了149.4%，在越南语客户端上，平均提升了322.3%。进一步地，FedEvent通过引入跨语言词嵌入，与未加入该模块（即$w/o\ NC$）的方法相比，在中文客户端上，平均提升了2.4%，在越南语客户端上，平均提升了4.3%。由于引入跨语言词嵌入模块后，FedEvent确保了不同客户端上输入特征空间的一致性，从而增强了蒸馏效果。可以发现，跨语言词嵌入模块（$\mathcal{M}$）在越南语客户端上表现尤为突出，可能是mBERT对越南语的表征能力不如对中文那样强大，利用$\mathcal{M}$对消息进行表征后，效果更加明显。

此外，在各个客户端上，相较于未加入锚点对齐损失（即$w/o\ NA$）的方法，FedEvent在NMI和AMI上平均下降了5.3%，进一步印证了锚点对齐损失的重要性。在缺

少锚点对齐损失的情况下，低资源客户端不易捕获到锚点对齐的信息，导致蒸馏效果较差。

### 3.7 案例分析

为了更直观地评估本文所提出方法的蒸馏效果，我们在英文和越南语的数据集$M_{17}$块上进行了案例分析。其中，英文客户端$M_{17}$块上共有2676条消息，分散在35个簇中。越南语客户端$M_{17}$块上共有350条消息，分散在14个簇中。具体来说，我们利用混淆矩阵来探究我们的方法以及KPGNN方法（单机）的效果，从而更直观地分析FedEvent获取了多少先验知识。

混淆矩阵是一种特定的表格布局，用于可视化监督学习算法的性能。在混淆矩阵中，每一行代表实际类别，每一列代表预测类别。矩阵的每个单元格包含了在该实际类别和预测类别下的样本数量，可以更直观的了解预测值与真实值之间的差异。本实验中，混淆矩阵的行数和列数与越南语客户端的类簇个数一致。考虑到基于密度的聚类方法所获得的事件数量与实际事件数量之间的差异，我们统一采用K-Means方法来获得预测标签。本实验共构建了三个混淆矩阵：第一个混淆矩阵（$FedEvent_{en}$）代表本文方法中富资源客户端上的真实结果与预测结果之间的关系；第二个混淆矩阵（$FedEvent_{vi}$）代表低资源客户端上的真实结果与预测结果之间的关系；第三个混淆矩阵（$KPGNN_{vi}$）代表KPGNN算法在越南语消息块$M_{17}$块上的预测结果与真实结果之间的关系。

通过三个混淆矩阵，我们发现$FedEvent_{vi}$的聚类结果在对角线上大多显示出比$KPGNN_{vi}$更高的值，意味着其预测结果更为集中，准确度更高。这一发现表明，即便在低资源环境下，FedEvent也能有效地帮助模型获取到一定的先验知识，从而提升性能。相反，KPGNN由于数据的限制，其模型性能受到了显著影响。进一步地，我们观察到当富资源客户端在某一类簇上的表现会直接影响低资源客户端在这一类簇上性能表现，这说明富资源客户端会将边界知识蒸馏至低资源客户端。而对于KPGNN而言，由于缺乏足够的先验知识，它在面对相似文本或共享相同元素的文本时，往往会出现判别能力不足的情况，从而导致模型识别错误。

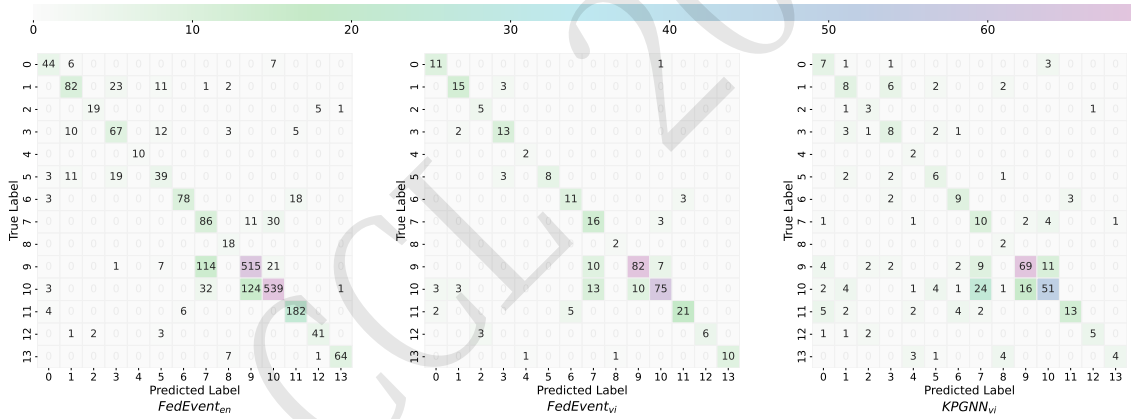



图 3. 消息块$M_{17}$混淆矩阵实验结果

### 3.8 可视化

为了直观地展示我们所提出的FedEvent框架的聚类效果，我们采用t-SNE(Van der Maaten and Hinton, 2008)对消息块$M_7$的表示进行了降维可视化。可视化结果如图4所示。与单机以及多机环境下的方法相比，FedEvent能够学习到更加可区分的表示。同时，观察到FedEvent比$w/o\ NA$聚集的更加紧密，验证了锚点对齐损失的重要性。此外，跨语言词嵌入模块的引入同样也提升了模型性能。

## 4 结论

本文提出了基于联邦蒸馏的跨语言社交媒体事件检测框架。该框架采用生命周期机制以适应在线及离线场景下的事件检测需求，并通过过程监督以及结果监督实现知识的有效迁移。本方法通过知识蒸馏技术，有效缓解了面对低资源语言时的挑战，并利用跨语言词嵌入模块，实现了不同语言间的语义空间映射。通过实验证明了我们所提出的方法的有效性。据我们所知，


(c) 2024 中国中文信息学会计算语言学专业委员会

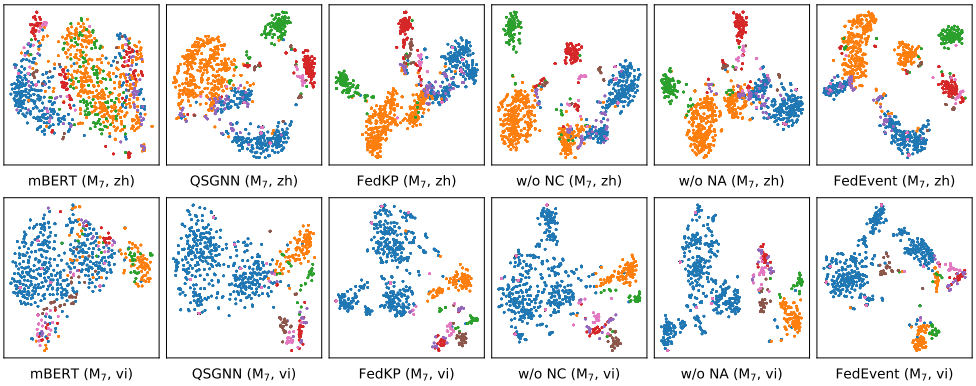



图 4. FedEvent可视化图

这是首个在联邦学习环境下，针对低资源语言环境进行社交媒体事件检测的研究。在下一步的研究工作中，我们计划进一步探索针对低资源场景下长尾数据分布的社交媒体事件检测策略。

## 参考文献

Charu C Aggarwal and Karthik Subbian. 2012. Event detection in social streams. In Proceedings of the 2012 SIAM international conference on data mining, pages 624–635. SIAM.

Farzindar Atefeh and Wael Khreich. 2015. A survey of techniques for event detection in twitter. Computational Intelligence, 31(1):132–164.

Hila Becker, Mor Naaman, and Luis Gravano. 2011. Beyond trending topics: Real-world event identification on twitter. In Proceedings of the international AAAI conference on web and social media, volume 5, pages 438–441.

Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. 2013. Density-based clustering based on hierarchical density estimates. In Jian Pei, Vincent S. Tseng, Longbing Cao, Hiroshi Motoda, and Guandong Xu, editors, Advances in Knowledge Discovery and Data Mining, pages 160–172, Berlin, Heidelberg. Springer Berlin Heidelberg.

Yuwei Cao, Hao Peng, Jia Wu, Yingtong Dou, Jianxin Li, and Philip S Yu. 2021. Knowledge-preserving incremental social event detection via heterogeneous gnns. In Proceedings of the Web Conference 2021, pages 3383–3395.

Xueqi Cheng, Xiaohui Yan, Yanyan Lan, and Jiafeng Guo. 2014. Btm: Topic modeling over short texts. IEEE Transactions on Knowledge and Data Engineering, 26(12):2928–2941.

Hyung Won Chung, Dan Garrette, Kiat Chuan Tan, and Jason Riesa. 2020. Improving multilingual models with language-clustered vocabularies. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 4536–4546, Online, November. Association for Computational Linguistics.

Mário Cordeiro. 2012. Twitter event detection: combining wavelet analysis and topic inference summarization. In Doctoral symposium on informatics engineering, volume 1, pages 11–16.

Wanqiu Cui, Junping Du, Dawei Wang, Feifei Kou, and Zhe Xue. 2021. Mvgan: Multi-view graph attention network for social event detection. ACM Transactions on Intelligent Systems and Technology (TIST), 12(3):1–24.

Rui Gaspar, Cláudia Pedro, Panos Panagiotopoulos, and Beate Seibt. 2016. Beyond positive or negative: Qualitative sentiment analysis of social media reactions to unexpected stressful events. Computers in Human Behavior, 56:179–191.

Linmei Hu, Bin Zhang, Lei Hou, and Juanzi Li. 2017. Adaptive online event detection in news streams. Knowledge-Based Systems, 138:105–112.

J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. ArXiv, abs/2106.09685.


(c) 2024 中国中文信息学会计算语言学专业委员会

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Bang Liu, Fred X Han, Di Niu, Linglong Kong, Kunfeng Lai, and Yu Xu. 2020a. Story forest: Extracting events and telling stories from breaking news. ACM Transactions on Knowledge Discovery from Data (TKDD), 14(3):1–28.

Fanzhen Liu, Shan Xue, Jia Wu, Chuan Zhou, Wenbin Hu, Cecile Paris, Surya Nepal, Jian Yang, and Philip S Yu. 2020b. Deep learning for community detection: progress, challenges and opportunities. arXiv preprint arXiv:2005.08225.

Yaopeng Liu, Hao Peng, Jianxin Li, Yangqiu Song, and Xiong Li. 2020c. Event detection and evolution in multi-lingual social streams. Frontiers of Computer Science, 14:1–15.

Fabrizio Marozzo and Alessandro Bessi. 2018. Analyzing polarization of social media users and news sites during political campaigns. Social Network Analysis and Mining, 8:1–13.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In Artificial intelligence and statistics, pages 1273–1282. PMLR.

Andrew J McMinn, Yashar Moshfeghi, and Joemon M Jose. 2013. Building a large-scale corpus for evaluating event detection on twitter. In Proceedings of the 22nd ACM international conference on Information & Knowledge Management, pages 409–418.

Tasnim Mohiuddin, M Saiful Bari, and Shafiq Joty. 2020. LNMap: Departures from isomorphic assumption in bilingual lexicon induction through non-linear mapping in latent space. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2712–2723, Online, November. Association for Computational Linguistics.

Duc T Nguyen and Jason J Jung. 2015. Real-time event detection on social data stream. Mobile Networks and Applications, 20:475–486.

Tahir M Nisar and Man Yeung. 2018. Twitter as a tool for forecasting stock market movements: A short-window event study. The journal of finance and data science, 4(2):101–119.

Ozer Ozdikis, Pinar Karagoz, and Halit Oğuztüzün. 2017. Incremental clustering with vector expansion for online event detection in microblogs. Social Network Analysis and Mining, 7:1–17.

Hao Peng, Jianxin Li, Qiran Gong, Yangqiu Song, Yuanxing Ning, Kunfeng Lai, and Philip S Yu. 2019. Fine-grained event categorization with heterogeneous graph convolutional networks. arXiv preprint arXiv:1906.04580.

Hao Peng, Jianxin Li, Yangqiu Song, Renyu Yang, Rajiv Ranjan, Philip S Yu, and Lifang He. 2021. Streaming social event detection and evolution discovery in heterogeneous information networks. ACM Transactions on Knowledge Discovery from Data (TKDD), 15(5):1–33.

Hao Peng, Ruitong Zhang, Shaoning Li, Yuwei Cao, Shirui Pan, and Philip S. Yu. 2023. Reinforced, incremental and cross-lingual event detection from social messages. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(1):980–998.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In Anna Korhonen, David Traum, and Lluís Màrquez, editors, Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4996–5001, Florence, Italy, July. Association for Computational Linguistics.

Jiaqian Ren, Lei Jiang, Hao Peng, Yuwei Cao, Jia Wu, Philip S Yu, and Lifang He. 2022a. From known to unknown: quality-aware self-improving graph neural network for open set social event detection. In Proceedings of the 31st ACM International Conference on Information & Knowledge Management, pages 1696–1705.

Jiaqian Ren, Lei Jiang, Hao Peng, Zhiwei Liu, Jia Wu, and S Yu Philip. 2022b. Evidential temporal-aware graph-based social event detection via dempster-shafer theory. In 2022 IEEE International Conference on Web Services (ICWS), pages 331–336. IEEE.

Jiaqian Ren, Hao Peng, Lei Jiang, Jia Wu, Yongxin Tong, Lihong Wang, Xu Bai, Bo Wang, and Qiang Yang. 2023. Transferring knowledge distillation for multilingual social event detection.


(c) 2024 中国中文信息学会计算语言学专业委员会

Xing Su, Shan Xue, Fanzhen Liu, Jia Wu, Jian Yang, Chuan Zhou, Wenbin Hu, Cecile Paris, Surya Nepal, Di Jin, et al. 2022. A comprehensive survey on community detection with deep learning. IEEE Transactions on Neural Networks and Learning Systems.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. Journal of machine learning research, 9(11).

Chuhan Wu, Fangzhao Wu, Lingjuan Lyu, Yongfeng Huang, and Xing Xie. 2022. Communication-efficient federated learning via knowledge distillation. Nature communications, 13(1):2032.

Weiren Yu, Jianxin Li, Md Zakirul Alam Bhuiyan, Richong Zhang, and Jinpeng Huai. 2017. Ring: Real-time emerging anomaly monitoring system over text streams. IEEE Transactions on Big Data, 5(4):506–519.

Kuo Zhang, Juan Zi, and Li Gang Wu. 2007. New event detection based on indexing-tree and named entity. In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pages 215–222.

Xiangmin Zhou and Lei Chen. 2014. Event detection over twitter social media streams. The VLDB journal, 23(3):381–400.

CCL 2024

# 基于生成式语言模型的立场检测探究

**张袁硕[1,2], 李澳华[1,2], 尹召宁[1,2], 王潘怡[1], 陈波[1,2,3,*], 赵小兵[1,2,3,*]**
1.中央民族大学,信息工程学院,北京,100081
2.国家语言资源监测与研究少数民族语言中心
3.国家安全研究院语言信息安全研究中心
chenbomuc@muc.edu.cn

## 摘要

近年来，立场检测任务受到越来越多的关注，但相关标注数据在范围和规模上都有限，不能有效支撑基于神经网络的立场检测。为此，本文探索在零样本/少样本场景下生成式语言模型在立场检测任务上的能力。首先，构建了一个全新的面向立场检测的数据集，包含5个主题，共2500个人工标注样例；然后，在此数据集上进行了一系列探索实验，实验结果表明：生成式语言模型在零样本设定下，采用结构化的提示学习表现良好；增加额外信息能够显著提升模型性能；在少样本设定下，提供相同目标的示例能够明显提升模型性能，而不同目标示例产生了负面作用；使用思维链可以显著提升模型性能；受提示学习的启发，微调预训练语言模型进一步论证提供额外信息对立场检测的增益显著。

**关键词：** 生成式语言模型 ； 立场检测 ； 零样本/少样本 ； 提示学习

# Research on Stance Detection with Generative Language Model

**Yuanshuo Zhang[1,2], Aohua Li[1,2], Zhaoning Yin[1,2], Panyi Wang[1,2]**
**Bo Chen[1,2,3,*], Xiaobing Zhao[1,2,3,*]**
1. School of Information Engineering, Minzu University of China
Beijing 100081, China
2. National Language Resource Monitoring and Research Center of Minority Languages
3. Language Information Security Research Center Institute of National Security MUC
chenbomuc@muc.edu.cn

## Abstract

In recent years, stance detection has received increasing attention, but the relevant annotated data is limited in scope and scale, which cannot effectively support neural network-based stance detection methods. Therefore, this paper explores the ability of generative language models in stance detection in zero/few shot settings.Firstly, a new stance detection dataset is constructed, which contains 5 topics and a total of 2500 manually labelled samples; then, a series of exploratory experiments are carried out on this dataset, and the experimental results show that: the generative language model performs well in structured prompt learning under zero-shot setting; fusion of extra information can significantly improve model performance; under the setting of few-shot, providing examples with the same goal can significantly improve model performance, while examples with different goals have a negative impact; Chain-of-Thought can significantly improve model performance; Inspired by prompt learning experiments, this paper fine tunes the pre-trained language model to further demonstrate that providing additional information has a significant gain in stance detection.

**Keywords:** Generative Language Model , Stance Detection , Zero/Few Shot , Prompt Learning


(c) 2024 中国中文信息学会计算语言学专业委员会

## 1 引言

随着社交媒体的普及和信息时代的到来，人们在各种在线平台上表达对热门话题的不同观点的意愿上升，因此，产生了大量的用户生成内容。基于这些生成内容，相关组织或机构可以准确掌握公众对特定话题的态度和立场，为相关决策提供依据。面向此类应用情境的立场检测 (李洋et al., 2021)任务也因此吸引了广泛的关注，该任务目标是给定文本，判定其作者对特定话题、观点或事件的立场或态度（支持、反对或者中立）。例如：针对"深圳禁摩限电"这一政策法规，用户发表微博"广州的也给全部禁了吧，特别是摩托车，容易出事！" (Xu et al., 2016b)，从"全部禁了"、"出事"这类词语表达用户对该目标的支持立场。

立场检测是自然语言处理领域的一个重要研究任务，一般将其建模为分类问题。目前，立场检测主流方法是深度学习 (Igarashi et al., 2016)的方法。尽管现有的立场检测模型通过有监督学习可以在一些常见的任务设定，如单目标 (Mohammad et al., 2016)、多目标 (Sobhani et al., 2017)等设定上表现出色，但在少样本（Few-shot）和零样本（Zero-shot） (Allaway and McKeown, 2020)设定下，即模型在没有或只有少量针对特定话题的标注数据的情况下，其性能显得不尽人意。

随着生成式语言模型 (Radford and Narasimhan, 2018)的兴起，如GPT-3，因其在预训练阶段通过大规模无监督数据的学习，使其掌握了丰富的语言知识，具有强大的语言理解和生成能力，同时也具备很强的任务泛化能力，在新任务上表现出色。在零样本设定下采用合适的提示就能取得不错的效果 (Wu et al., 2021)，在少样本设定下采用示例或思维链（Chain of Thought，COT）的提示方式能进一步提升效果 (Wei et al., 2022)。Floridi等人(2020)讨论了GPT-3的性质、范围、限制和影响，并探讨了其在数学、语义（图灵测试）和伦理问题上的能力；Ye等人(2023)使用21个数据集评估了GPT-3和GPT-3.5在九个自然语言理解任务上比较了零样本和少样本场景下每个任务的表现。

在本文中，我们针对立场检测任务，探索基于生成式语言模型的立场判别方法。我们先构建了一个新的人工标注的立场检测数据集，然后在此数据集上探究生成式语言模型的立场判别能力，特别是在零样本和少样本设定下，并尝试回答4个问题：Q1：生成式语言模型在零样本立场检测问题中的效果如何？Q2：补充额外的信息能否提升模型的表现？Q3：在少样本设定下，示例能否提升模型的表现？Q4：思维链能否提升模型的表现？

本文贡献如下：

- 针对立场检测任务，我们构建了一个新的数据集，该数据集包含来自微博的5个热点话题，每个话题收集并人工标注了500条用户生成内容，可作为中文立场检测单目标、多目标、零样本和少样本等不同设定下的标准数据集。
- 我们从多个角度对生成式语言模型基于提示的立场判别进行了探索，尝试在提示中增加不同的额外知识信息，并采用多样化的方式，以实验出面向生成式语言模型最有效的提示方法。
- 我们对预训练语言模型进行微调，进一步证实了明确的立场标签和事件背景信息对立场检测有重要的影响。

## 2 相关工作

### 2.1 立场检测

近年来，立场检测受到广泛的关注，并得到了应用，如舆情分析 (李洋et al., 2021)、社交媒体监测 (刘高勇et al., 2022)和政治舆论分析 (Thomas et al., 2006)等。立场检测方法的发展经历了基于传统机器学习的方法，到基于神经网络的方法 (Igarashi et al., 2016)，再到基于大规模预训练语言模型的方法的演变。基于传统机器学习的方法，主要依赖于手工定义特征，常用特征包括：文本特征、情感特征和混合特征 (奠雨洁et al., 2017)等，再利用分类模型进行立场分类，如支持向量机 (Xu et al., 2016a)。神经网络方法兴起后，在立场检测任务中，为了减少了人工构造特征的工作量，神经网络被引入来自动学习更丰富的特征。Mohtarami等人(2018)提

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

出了基于端到端的记忆网络模型。杨顺成等人(2020)提出了基于GCN和Bi-LSTM的微博立场检测方法。面向零样本/少样本情境，预训练语言模型可以从大规模的无标注语料中学习到丰富且高层次的语言知识，能很好地作用于下游任务，并具有一定的泛化能力。Hosseinia等人(2020)提出了一种基于BERT的问题表示与情绪信息相结合的立场检测的方法。Allaway等人(2020)基于BERT和Attention机制开发了Topic-Grouped Attention(TGA)模型，通过上下文聚类获得的广义主题表示来利用有关主题相似性的信息。He等人(2022)利用维基百科中关于目标的背景知识来增强立场检测，提出了具有两种变体的WS-BERT来编码此类知识。Li等人(2023a)提出了知识增强立场检测框架(KASD)，在社交媒体立场检测任务中引入了情景知识和话语知识，并利用ChatGPT对上述两种知识进行提取和注入，性能显著提高。Li等人(2023b)提出了目标立场提取（Target-Stance Extraction, TSE）任务，这种新任务有助于促进立场检测领域的未来研究。

### 2.2 生成式语言模型的零样本和少样本学习能力探究

生成式语言模型拥有优秀的任务泛化能力，可直接应用于各种新任务。在零样本方面,Kojima等人(2022)提出了一种基于零样本学习的思维链提示，在各种推理任务上的表现明显优于零样本大语言模型的性能；李燕等(2023)将ChatGPT作为辅助工具，利用提示和模型先验知识助力银行内部的渗透测试；卢宇等(2023)探究了ChatGPT在教育方面的题目生成、自动结题和辅助批阅的能力；寿建琪(2023)使用GPT助力信息检索，提出了一种结合OPAC类检索服务和基于LLM的智能化信息检索服务的自适应文献检索框架。在少样本方面，Brown等(2020)通过预训练大型语言模型并使用少量样本进行微调，在多种NLP任务上都表现出色，例如翻译、问答和完形填空任务，展示了GPT-3强大的少样本学习能力。冯广敬等人(2022)提出了基于DialoGPT的二阶段对话生成模型；马志强等人(2023)基于GPT-2构建情感导向对话回复生成实验系统EGRGM；言佳润等人(2023)在GPT、BERT、RoBERTa预训练模型上进行提示学习，通过P-tuning自动构建连续模板进行论辩挖掘。Tuckute等人(2024)基于GPT的编码模型来预测大脑对多样化句子的反应，并利用这些模型选择的句子在新个体中记录大脑反应验证模型的预测能力。Wachowiak等人(2023)探究了GPT-3 检测隐喻语言和预测隐喻源域的能力。Black等人(2022)介绍了一个在Pile 上训练的200亿参数自回归语言模型GPT-NeoX-20B，一个特别强大的少样本推理器。我们首次将生成式语言模型应用于立场检测，并通过系列对比实验探索最有效的提示方法。

## 3 生成式立场检测

为了探究生成式语言模型在立场检测任务上的效果，特别是在零样本和少样本的设定。我们设计了一系列实验，旨在回答Q1-Q4。

实验1：零样本设定下，使用常规提示和结构化提示，旨在探索最适合立场检测任务的提示，并分析在零样本设定下生成式语言模型立场检测的性能。（跟Q1有关）

实验2：在提示中加入额外的信息，包括：事件相关的背景和明确的立场标签。例如，针对社会热点事件“恶意殴打他人者妻女被网暴”，有关事件的背景知识：“唐山男子烧烤店无缘无故殴打女子，事后男子妻女遭到网暴。”和明确的立场标签：“活该被网暴（支持）、不该被网暴（反对）、中立。”实验2旨在探究通过补充额外的信息能否提升生成式语言模型在立场检测任务上的性能。（跟Q2有关）

实验3：少样本设定下，在提示中加入立场检测示例。实验3主要考察生成式语言模型在给定相同或者不同目标的示例后，能否提升立场判别的准确度。（跟Q3有关）

实验4：在提示信息中引入思维链，旨在验证引入推理过程后，能否提升立场判别的准确度。（跟Q4有关）

## 4 实验及分析

为了探索生成式语言模型在立场检测任务上的效果，我们先构建了一个新的立场检测数据集，接着选择代表性的生成式语言模型进行实验1-4，最后对实验结果进行详细分析。我们使用F1值作为立场检测的评价指标。对于没有给出明确立场的回答，我们将其视为判断错误。


(c) 2024 中国中文信息学会计算语言学专业委员会

### 4.1 微博立场检测数据集

我们从新浪微博平台选择了近期发生的5个具有争议的话题，包括“恶意殴打他人者妻女被网暴”、“女子不让6岁男童上女厕所遭痛骂”、“警方通告胡鑫宇为自杀”、“满江红起诉大V”和“泼水节女生选择原谅对方”，并将这5个话题作为立场检测的目标；然后，我们检索了5个话题的背景知识并爬取了对应评论，并利用启发式规则从目标中提取了明确的立场标签；最后，我们对评论进行立场标注，在标注过程中，采用多人交叉标注的方式，以确保数据标注的客观性和准确性。具体而言，三位标注员（均为NLP领域硕士研究生）对每一条评论文本根据目标话题进行独立的立场判断和标注。收集了三份独立的标注数据后，进行比对，若某条评论文本得到两人以上标注的结果相同，则该立场标注被视为正确标注。如果三位标注员对某条评论文本的立场标注存在不一致，即标注了三个不同的立场，那么该评论文本将被排除在数据集之外。

我们数据集的优势主要有两点：1. 数据集更具实用性，数据集全部针对微博平台的热点舆情事件，在该数据集上的研究能够快速对接到真实的微博平台立场分析的应用场景；2.数据集更具有挑战性，数据集中的评论针对的是一个复杂事件，包含多个子事件，如“恶意暴打他人者妻女被网暴”，包含了男子殴打女子，男子的妻女被网友网暴，要对评论给出准确的立场判定，需要有足够的事件知识信息和严谨的推理。

我们将该数据集命名为Weibo-SD并公开共享于github[0]。该数据集共有5个目标，每个目标有500个评论，每个目标下的背景、明确的立场标签和评论的立场分布如表1所示。

| 主题 | 明确的立场标签 | 背景知识 | 支持 | 反对 | 中立 |
|---|---|---|---|---|---|
| 恶意殴打他人者妻女被网暴 | 活该被网暴(支持),不该被网暴(反对),中立 | 唐山男子烧烤店无缘无故殴打女子，事后男子妻女遭到网暴。 | 260 | 62 | 178 |
| 女子不让6岁男童上女厕所遭痛骂 | 活该被骂(支持),不该被骂(反对),中立 | 女子不让6岁男童上女厕所，遭男童妈妈当众辱骂。 | 34 | 338 | 128 |
| 警方通告胡鑫宇为自杀 | 认为是自杀(支持),怀疑不是自杀(反对),中立 | 高一男生胡鑫宇傍晚从宿舍至教学楼路上离奇失踪，多点暗示胡鑫宇有自杀倾向。最终胡鑫宇在学校后山被找到，警方通报其为自缢身亡。 | 127 | 226 | 247 |
| 满江红起诉大V | 应该起诉(支持),不该起诉(反对),中立 | 2023年春节电影《满江红》被指控“幽灵场”“抄袭”等。随后满江红官方发文，对造谣者沈逸、屠龙的胭脂井、平原公子赵胜、喵斯拉大王二号机四位大V提出诉讼。 | 61 | 195 | 224 |
| 泼水节女生选择原谅对方 | 应该原谅(支持),不该原谅(反对),中立 | 泼水节女子遭众男子围着泼水撕雨衣，事后女子选择原谅对方，不立案。 | 147 | 176 | 177 |

Table 1: Weibo-SD数据集

### 4.2 生成式语言模型

针对要探究的问题，我们选择了2个生成式语言模型（在本文中生成式语言模型与大语言模型（Large Language Model）两种表述是通用的，并采用LLM简称），分别是GPT-3.5-turbo和ChatGLM2-6B，分别简称为GPT和GLM。所有实验均采用API调用的方式完成，每个结果都是5次实验结果的平均值。

### 4.3 实验结果及分析

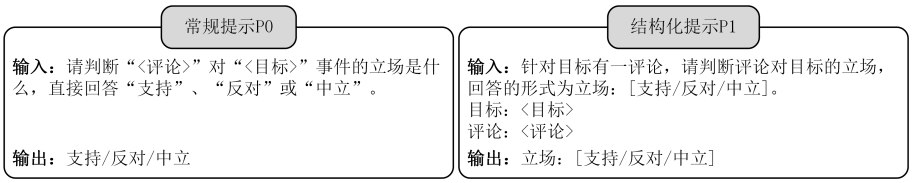



Figure 1: 提示模板P0、P1

[0] https://anonymous.4open.science/r/Weibo-SD-618D/


(c) 2024 中国中文信息学会计算语言学专业委员会

| 提示 | GPT | GLM |
|---|---|---|
| P0 | 35.28 | 35.12 |
| P1 | **42.68** | **47.52** |

Table 2: 在零样本设定下使用两种提示进行提问(单位%)

针对实验1：在零样本设定下，采用常规提示P0和结构化提示P1两种提示模板，提示模板如图1所示，立场检测的结果如表2所示。可以看到，使用常规提示时，GPT和GLM的F1值均在35.20左右。使用结构化提示时，GPT的F1从35.28提升到42.68，提升了7.40；GLM的F1从35.12提升到47.52，提升了12.40。GPT和GLM的效果均有显著提升，且GLM的效果优于GPT4.84个百分点。我们认为使用结构化提示较常规提示能显著提升LLM在立场检测任务上的性能。结构化提示可以引导模型关注于文本中的关键信息，帮助模型更好地理解任务需求，从而提高立场判别的准确性。因此，在后续的实验中我们均使用结构化提示进行提问。GLM（ChatGLM2-6B）发布于2023年，预训练语料中可能已经包含2022年的数据，而用于预训练GPT的数据是基于2021年之前的数据，因此在零样本且不提供额外信息的情况下，GLM的效果要优于GPT。

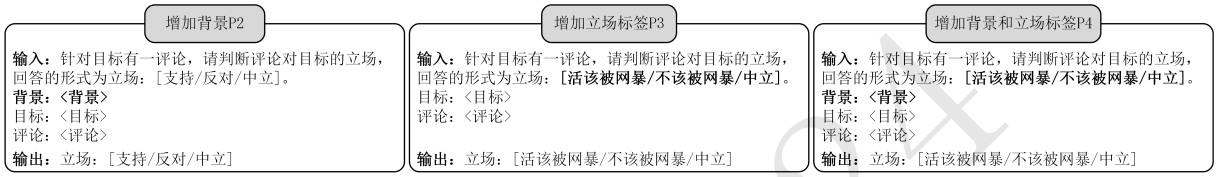



Figure 2: 提示模板P2、P3、P4

| 加入额外信息 | 提示 | GPT | GLM |
|---|---|---|---|
| 无额外信息 | P1 | 42.68 | **47.52** |
| 背景 | P2 | 44.20 | 47.04 |
| 立场标签 | P3 | 50.32 | 40.52 |
| 背景+立场标签 | P4 | **51.44** | 40.48 |

Table 3: 在提示中增加背景和明确的立场标签(单位%)

针对实验2：我们尝试在提示中增加了额外的信息，包括背景知识和明确的立场标签。提示模板如图2所示，实验结果表3所示。

对于GPT，在提升中增加额外信息对立场判别效果提升显著。具体来说，相比无额外信息的提示模板P1，增加了背景信息后F1值为44.20，提升了1.52；增加了明确立场标签后F1值为50.32，提升了7.64，提升显著；同时增加了背景和明确立场标签后F1值达到51.44，提升了8.76，取得了最好的效果。我们认为提供额外的背景知识能够帮助模型更深入地理解文本的上下文，从而提高立场检测的准确性；我们的立场检测数据集包含的都是复杂的话题，话题本身可能包含2个事件，如“恶意殴打他人者妻女被网暴”中就包含“某男子恶意殴打他人”（事件A）和“该男子的妻女被网暴”（事件B）两个事件，简单的立场标签，如“支持”、“反对”和“中立”，在零样本或少样本的情况下，很难让LLM学习到支持的立场到底是支持事件A还是B，特别是事件A和B往往是对立的情况下。明确的立场标签可以告诉模型支持的目标是什么，反对目标是什么，从而消除模型的混乱，提高立场检测性能。例如对于“恶意殴打他人者妻女被网暴”事件，支持的立场等同于“活该被网暴”，而不是支持打人者或者打人者的妻女。

对于GLM，增加了背景信息后F1有小幅下降，下降了0.48；增加了明确立场标签后F1大幅下降，从47.52下降到40.52，下降了7.00；同时增加了背景和明确立场标签和只增加明确立场标签F1值接近，下降了7.04。我们认为GLM的训练语料已经包含这些事件相关数据，因此增加背景信息对GLM影响不大；观察输出结果，我们发现GLM会因明确立场标签中“网暴”、“被骂”等敏感词汇产生误判，认为“网暴”、“骂人”是错误的行为，将立场判定为“不该网暴”、“不


(c) 2024 中国中文信息学会计算语言学专业委员会

该被骂”或保持中立。我们统计了去除包含敏感词汇立场后的F1值，分别为45.19和45.13，证实了我们的猜想是正确的。

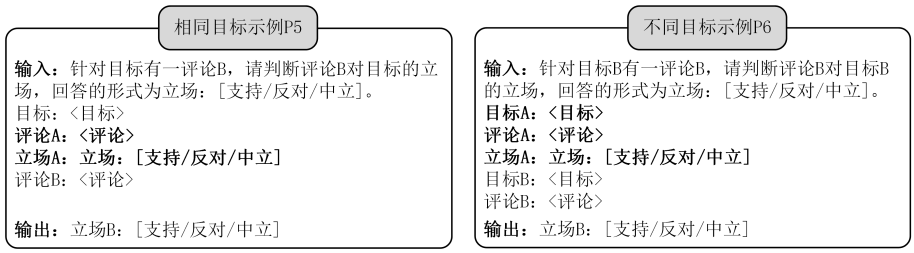



Figure 3: 提示模板P5、P6

| 示例 | 提示 | GPT | GLM |
|---|---|---|---|
| 无示例 | P1 | 42.68 | **47.52** |
| 相同目标示例 | P5 | **47.92** | 43.12 |
| 不同目标示例 | P6 | 42.56 | 42.58 |

Table 4: 在少样本设定下在提示中添加示例(单位%)

针对实验3：探究少样本设定下，在提示中加入示例对立场判别的影响。提示模板如图3所示，结果如表4所示。

对于GPT，添加相同目标示例后F1有明显提升，从42.68提升到了47.92，提升了5.24；添加不同目标示例后F1小幅下降，从42.68下降到了42.56，下降了0.12。我们认为提供相同目标的示例能使模型更准确地捕捉到与指定目标特定立场相关的词汇、短语和句子结构，从而提高检测的准确性，而不同目标的示例可能会带来干扰，导致模型性能下降。

对于GLM，添加相同目标示例或不同目标示例后，模型的性能均有明显降低，平均降低了4.67。观察输出结果后，我们发现GLM有时会出现将示例评论与待检测评论混淆或将示例目标与待检测评论目标混淆的问题，例如忽略了评论B而回答示例评论A的立场，或将示例目标A当作待检测目标。我们认为加入示例会使提示复杂化，GLM不能完全理解示例的意图，导致性能下降。

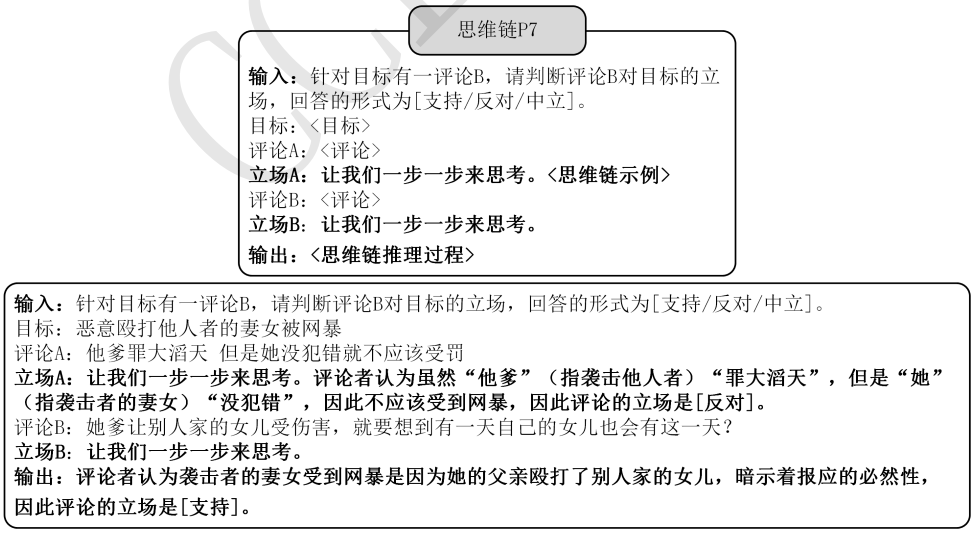



Figure 4: 提示模板P7及示例

针对实验4：探究引入思维链后，能否提升立场判别的准确度。为了更直观的展示思维链提示的过程，我们提供了一个示例。示例和提示模板如图4所示，结果如表5所示。

| 提问方式 | 提示 | GPT | GLM |
|---|---|---|---|
| 无思维链 | P1 | 42.68 | **47.52** |
| 思维链 | P7 | **51.36** | 28.84 |

Table 5: 在提示中添加思维链(单位%)

对于GPT，引入思维链后F1值显著提升，从42.68提升到了51.36，提升了8.68。我们认为，相比于直接给出评论的立场，思维链提供了中间的推导过程，减少了跳跃性推理，使模型不只关注于输入评论的表面特征，还能够捕捉到评论背后的复杂关系，如评论中的人物关系、与人物相关的信息等。例如评论“她爹罪大滔天，但她没有犯错就不该受到惩罚”，评论者认为虽然“他爹”（指袭击他人者）“罪大滔天”，但是“她”（指袭击者的妻女）“没犯错”，因此不应该受到惩罚，这表明评论者对袭击者的妻女感到同情并认为她们不应该因为袭击者的行为而受到处罚，因此评论的立场是[反对]。

对于GLM，引入思维链后F1值大幅降低，降低了18.68。观察输出结果，发现有近20%的输出答非所问，不能输出明确的立场。我们认为原因有两点，一是GLM的参数规模只有60亿,本身的上下文理解能力和推理能力有限；二是思维链的引入可能会使模型在处理问题时过度复杂化，使本就能力受限的模型更难以推理。

#### 4.3.1 立场判别偏向性

我们对LLM的立场判别的偏向性进行了评估。实验结果如表6所示。可以发现，GPT对于中立立场的判别准确度整体很高，基于提示P1的准确度达到了最高的83.52，平均值也高达71.61，而对支持判定的准确度较低，平均值分别为27.92。在加入了明确立场标签的P3、P4和引入了思维链的P7中，GPT对反对立场的判别明显高于其他提示模板，分别为49.65、50.65、40.02，远高于平均值。我们认为GPT对中立立场的偏向性较高，当它不能确定评论的立场时，可能会采取保守的策略而选择中立，明确的立场标签和思维链都能明显提高模型的推理能力。而GLM对反对立场的准确度更高，基于P6的准确度达到了最高的83.55，平均值为49.06，但对于支持和中立的准确度较低，分别为33.09和36.27。我们认为GLM对一些中文词汇敏感度较高，当发现提示中出现“网暴”之类词汇时，倾向于将立场判定为反对，因此对反对的偏向性较高。另外，LLM整体对支持立场的判别的准确度不高，平均值在35%左右，但增加了额外信息或引入思维链后，这一现象有所改善。

| 提示 | 支持 | 反对 | 中立 |
|---|---|---|---|
| P0 | 15.42/38.63 | 5.52/54.36 | 83.52/10.64 |
| P1 | 36.72/19.08 | 15.35/80.44 | 78.15/30.43 |
| P2 | 41.65/22.26 | 20.96/70.61 | 72.54/37.99 |
| P3 | 32.75/42.45 | 49.65/20.16 | 63.73/62.36 |
| P4 | 33.70/42.29 | 50.65/20.16 | 65.10/62.36 |
| P5 | 53.42/46.10 | 26.28/39.92 | 68.65/44.62 |
| P6 | 43.24/14.79 | 14.94/83.55 | 73.57/13.96 |
| P7 | 46.74/39.11 | 40.02/23.27 | 67.62/27.80 |
| AVG | 37.96/33.09 | 27.92/49.06 | 71.61/36.27 |

Table 6: 不同提示下检测不同立场的准确度(GPT/GLM(单位%))

#### 4.3.2 主题的差异性

我们评估了LLM在各个主题上立场判别差异性。结果如表7所示。可以发现，“警方通告胡鑫宇为自杀、“满江红起诉大V”和“泼水节女生选择原谅对方”这三个主题的F1值整体较高，均在45%左右；而对于“恶意殴打他人者妻女被网暴”和“女子不让6岁男童上女厕所遭痛骂”主题的F1值较低。我们分析其原因有二：一是后三个主题的事件逻辑相对简单，只涉及到“胡鑫宇是否是自杀”、“满江红是否该起诉大V”以及“女生是否该原谅对方”，而其前两个主题基本都包含两个事件，如看似只包含1个事件的“女子不让6岁男童上女厕所遭痛骂”也包含“女子不让6岁


(c) 2024 中国中文信息学会计算语言学专业委员会

男童上女厕所”和“女子遭痛骂”这两个明显立场对立的事件；二是后三个主题下的用户评价多数立场鲜明，如“警方通告胡鑫宇为自杀”这个主题下的评论“就我而言，失踪106天截至到山上发现，且是自缢，这期间遇害地点和学校如此之近，警方进行地毯式搜索无果。现在突然间找到，没有自杀一说，我直接拍板，就是他杀。”这都有利于LLM对立场进行相对准确的判别。

| 提示 | 恶意殴打他人者妻女被网暴 | 女子不让6岁男童上女厕所遭痛骂 | 警方通告胡鑫宇为自杀 | 满江红起诉大V | 泼水节女生选择原谅对方 |
|---|---|---|---|---|---|
| P0 | 40.29/33.10 | 23.96/33.47 | 32.39/39.71 | 48.45/37.40 | 37.04/37.72 |
| P1 | 44.49/20.20 | 27.48/65.40 | 38.88/53.60 | 52.05/55.80 | 50.90/42.60 |
| P2 | 45.60/23.80 | 25.85/57.80 | 54.80/52.60 | 47.20/52.40 | 47.74/48.60 |
| P3 | 41.40/39.88 | 45.80/27.40 | 59.20/46.94 | 57.80/47.20 | 47.40/41.40 |
| P4 | 43.20/39.92 | 46.80/27.40 | 60.80/46.94 | 57.00/47.05 | 49.40/41.40 |
| P5 | 55.40/30.43 | 32.40/39.80 | 51.80/50.20 | 51.40/51.20 | 48.65/44.00 |
| P6 | 49.20/15.07 | 28.20/64.84 | 37.31/49.33 | 52.00/45.89 | 46.29/40.76 |
| P7 | 51.60/23.36 | 42.53/26.53 | 57.00/41.33 | 54.05/35.05 | 51.90/30.98 |
| AVG | 46.40/28.22 | 34.13/42.83 | 49.02/47.58 | 52.49/46.50 | 47.42/40.93 |

Table 7: 不同提示下五种主题的F1(GPT/GLM(单位%))

## 5 微调预训练语言模型

基于生成式语言模型的立场检测实验表明：1）明确的立场标签具有很大的影响；2）补充背景信息能增强立场检测。为了探究LLM与流行立场检测模型的性能差异并进一步论证上述结论，我们微调了预训练语言模型（BERT、RoBERTa和mT5），在基础的设置中只输入目标和评论（P1），对比设置中，分别加入事件背景（P2）、明确的立场标签（P3），还有组合加入的设置（P4）。我们将Weibo-SD数据集中的80%作为微调数据集，剩余的20%作为测试集。实验结果如表8所示。

| 输入 | BERT | RoBERTa | mT5 | GPT | GLM |
|---|---|---|---|---|---|
| 目标和评论（P1） | 63.6 | 61.4 | 60.0 | 42.68 | **47.52** |
| +事件背景（P2） | 64.8 | 62.2 | 60.2 | 44.20 | 47.04 |
| +明确标签（P3） | 65.4 | 62.4 | 60.8 | 50.32 | 40.52 |
| +明确标签和事件背景（P4） | **66.6** | **65.4** | **62.6** | **51.44** | 40.48 |

Table 8: 微调预训练语言模型F1(单位%)

我们发现:

(1)微调BERT和RoBERTa的结果总是优于微调mT5。我们认为，mT5在处理生成类任务时表现出色，但在分类任务上可能不如专注于理解文本深层含义的BERT和RoBERTa。

(2)当同时融合明确的立场标签和事件背景信息时，微调BERT取得了最好的效果，F1值达到66.6。我们认为，明确的立场标签和事件背景信息缓解了复杂的评论-目标-事件-立场关系所带来的混乱，从而提高立场检测的效果。

(3)LLM的立场检测能力能达到微调小规模语言模型的60%到80%，与在特定任务的数据集上进行训练的模型相比还有一定差距。

(4)事件背景可以提供足够的有用信息，同时减少噪声的引入。例如评论：“故意撕雨衣就过分了，必须立案追究责任”，其中“撕雨衣”、“立案”的信息在主题中并未提及，而是出现在事件背景中，这些信息表达了评论者对“原谅”持反对态度。

(5)明确的立场标签能增强立场检测。明确的立场标签使得模型在训练过程中能够聚焦于与立场相关的关键特征，例如针对主题“泼水节女生选择原谅对方”，明确的立场标签使模型聚焦于“应该或不该原谅”。


(c) 2024 中国中文信息学会计算语言学专业委员会

## 6 总结

本文探究了生成式语言模型在立场检测任务中的表现。实验结果表明，在零样本设定下，使用结构化的提示学习取得了更好的效果，尤其是在提供额外信息时。在少样本设定下，模型的性能也能得到不错的提升，尤其在引入思维链后，但当给的学习样本与测试样本不属于相同目标时，效果出现小幅下降。另外，受限于模型的规模大小，当提示过于复杂时，模型的理解能力反而变差。额外的实验和分析还表明：生成式语言模型存在一定的立场判别倾向性；事件逻辑越简单立场判别越容易。最后，我们通过微调预训练语言模型（BERT、RoBERTa和mT5），发现生成式语言模型在立场检测任务中展现出一定的潜力，并进一步证实了事件背景和明确的立场标签能增强立场检测。

本文的发现揭示了生成式语言模型在立场检测中的潜在价值，也强调了在特定情境下其局限性。在未来，我们将1）进一步扩充数据集，并且在运用大模型做立场检测时，预先将数据中的敏感词处理成等同语义的非敏感词，以减少此因素对大模型的干扰；2）研究探索更有效的提示方式和更全面的信息融合及更好的信息融合机制，如简化提示，同时保留有效信息，或分步提示，以提升模型对复杂提示的理解能力，以进一步提升生成式语言模型在这一领域的应用价值。

## 致谢

感谢所有匿名审稿人的宝贵意见。本项研究成果受国家社会科学基金重大项目（22&ZD035）、国家语委重点项目（ZDI145-61）、中央民族大学项目(GRSCP202316, 2023QNYL22, 2024GJYY43)资助。

## 参考文献

Emily Allaway and Kathleen McKeown. 2020. Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8913–8931.

Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, Usvsn Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. GPT-NeoX-20B: An open-source autoregressive language model. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 95–136.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Luciano Floridi and Massimo Chiriatti. 2020. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30(4):681–694.

Zihao He, Negar Mokhberian, and Kristina Lerman. 2022. Infusing knowledge from Wikipedia to enhance stance detection. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 71–77.

Marjan Hosseinia, Eduard Dragut, and Arjun Mukherjee. 2020. Stance prediction for contemporary issues: Data and experiments. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 32–40.

Yuki Igarashi, Hiroya Komatsu, Sosuke Kobayashi, Naoaki Okazaki, and Kentaro Inui. 2016. Tohoku at SemEval-2016 task 6: Feature-based model versus convolutional neural network for stance detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 401–407.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of Neural Information Processing Systems*, volume 35, pages 22199–22213.


(c) 2024 中国中文信息学会计算语言学专业委员会

Ang Li, Bin Liang, Jingqian Zhao, Bowen Zhang, Min Yang, and Ruifeng Xu. 2023a. Stance detection on social media with background knowledge. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15703–15717.

Yingjie Li, Krishna Garg, and Cornelia Caragea. 2023b. A new direction in stance detection: Target-stance extraction in the wild. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10071–10085.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41.

Mitra Mohtarami, Ramy Baly, James Glass, Preslav Nakov, Lluís Màrquez, and Alessandro Moschitti. 2018. Automatic stance detection using end-to-end memory networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 767–776.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

Parinaz Sobhani, Diana Inkpen, and Xiaodan Zhu. 2017. A dataset for multi-target stance detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 551–557.

Matt Thomas, Bo Pang, and Lillian Lee. 2006. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 327–335.

Greta Tuckute, Aalok Sathe, Shashank Srikant, Maya Taliaferro, Mingye Wang, Martin Schrimpf, Kendrick Kay, and Evelina Fedorenko. 2024. Driving and suppressing the human language network using large language models. *Nature human behaviour*, 8(3):544–561.

Lennart Wachowiak and Dagmar Gromann. 2023. Does GPT-3 grasp metaphors? identifying metaphor mappings with generative language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1018–1032.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903.

Shaohua Wu, Xudong Zhao, Tong Yu, Rongguo Zhang, Chong Shen, Hongli Liu, Feng Li, Hong Zhu, Jiangang Luo, Liang Xu, and Xuanwei Zhang. 2021. Yuan 1.0: Large-scale pre-trained language model in zero-shot and few-shot learning. *arxiv:2110.04725[cs.CL,cs.AI]*.

Jiaming Xu, Suncong Zheng, Jing Shi, Yiqun Yao, and Bo Xu. 2016a. Ensemble of feature sets and classification methods for stance detection. In *Proceedings of Natural Language Understanding and Intelligent Applications*, pages 679–688.

Ruifeng Xu, Yu Zhou, Dongyin Wu, Lin Gui, Jiachen Du, and Yun Xue. 2016b. Overview of nlpcc shared task 4: Stance detection in chinese microblogs. In *Proceedings of the ICCPOL 2016*, pages 907–916.

Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, Jie Zhou, Siming Chen, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. A comprehensive capability analysis of gpt-3 and gpt-3.5 series models. *arxiv:2303.10420[cs.CL]*.

冯广敬, 刘箴, 刘婷婷, 许根, 庄寅, 王媛怡, and 柴艳杰. 2022. 基于情感变量的二阶段对话生成模型. 中文信息学报, 36:102–111.

刘高勇, 黄靖钊, and 艾丹祥. 2022. 融合立场检测和主题挖掘的突发公共事件网络舆情演化研究. 广东工业大学学报, 39:32–40+48.

卢宇, 余京蕾, 陈鹏鹤, and 李沐云. 2023. 生成式人工智能的教育应用与展望——以chatgpt系统为例. 中国远程教育, 43:24–31+51.

奠雨洁, 金琴, and 吴慧敏. 2017. 基于多文本特征融合的中文微博的立场检测. 计算机工程与应用, 53:77–84.

寿建琪. 2023. 走向"已知之未知": gpt大语言模型助力实现以人为本的信息检索. 农业图书情报学报, 35:16–26.


(c) 2024 中国中文信息学会计算语言学专业委员会

李洋, 孙宇晴, and 景维鹏. 2021. 文本立场检测综述. 计算机研究与发展, 58:2538–2557.

李燕, 赖胜枢, and 蒋泽宁. 2023. Chatgpt在银行业信息安全建设的应用. 金融科技时代, 31:14–18.

杨顺成, 李彦, and 赵其峰. 2020. 基于gcn和bi-lstm的微博立场检测方法. 重庆理工大学学报*(*自然科学*)*, 34:167–173.

言佳润and 鲜于波. 2023. 面向中文网络对话文本的论辩挖掘——基于微调与提示学习的大模型算法. 中文信息学报, 37:139–148.

马志强, 王春喻, 贾文超, and 杜宝祥. 2023. 情感导向对话回复生成模型. 中文信息学报, 37:104–114.

CCL 2024

# 基于双图注意力网络的篇章级散文情绪变化分析方法

**李爱琳[1]，李旸[2]，王素格[1,3,*]，李书琪[1]**
1.山西大学计算机与信息技术学院，山西省 太原 030006
2.山西财经大学金融学院，山西省 太原 030006
3.山西大学计算智能与中文信息处理教育部重点实验室，山西省 太原 030006
wsg@sxu.edu.cn

## 摘要

在散文中，作者的情绪会伴随着文章的段落或者句子发生变化，比如从悲伤到快乐、从喜悦到愤怒。为此，本文构建散文情绪变化数据集，提出一种基于双图注意力网络的多种知识融合的情绪变化分析方法。首先，引入意象知识库，建立融合意象知识的句子表示；其次，构建上下文带权依赖图和语篇带权依赖图，通过融合上下文知识和语篇结构，建立了融合上下文知识、语篇结构的句子表示；同时设计愉悦效价识别层，获得融合愉悦效价信息的句子表示；在此基础上，将以上三者表示进行拼接，通过全连接网络得到最终的情绪变化结果。实验结果表明，本文提出的方法可以有效识别情绪变化，为散文阅读理解中的思想情绪变化类问题的解答提供帮助。

**关键词：** 情绪变化 ； 散文 ； 图注意力网络

# A Document-Level Emotion Change Analysis Method Based on DualGATs for Prose

**Ailin Li[1]，Yang Li[2]，Suge Wang[1,3,*]，Shuqi Li[1]**
1.School of Computer and Information Technology,
Shanxi University,Shanxi 030006
2.Shanxi University of Finance and Economics,Shanxi 030006
3.Key Laboratory Computational Intelligence and Chinese Information
Processing of Ministry of Education, Shanxi University,Shanxi 030006
wsg@sxu.edu.cn

## Abstract

In Prose, the author's emotions evolve within passages or sentences,transitioning from sadness to happiness, or from joy to anger. To this end, this paper constructs a dataset of prose emotion changes and proposes emotion analysis method based on multi-knowledge fusion utilizing Dual-Graph Attention Network. Firstly, this paper incorporates image knowledge to get the deep representation of sentence. Secondly, the context-weighted dependency graph and the discourse-weighted dependency graph are constructed, and the sentence representation is established by integrating context knowledge and discourse structure. Additionally, a pleasure valence recognition layer is designed to integrate pleasure valence information into the sentence representation. Experimental results indicate the method is effective for identifying emotion changes, which can provide help for understanding the ideological and emotional changes in prose comprehension.

**Keywords:** Emotion change , Prose , Graph Attention Network

第二十三届中国计算语言学大会论文集，第492页-第503页，太原，中国，2024年7月25日至28日。
卷1：主会论文
(c) 2024 中国中文信息学会计算语言学专业委员会

## 1 引言

散文在文学作品中的表现形式较为自由和灵活(王小环, 2023)，并具有较强的抒情性。散文中的情绪变化是指在文学作品中，作者通过语言表达和叙述，展现了自身情绪的波动、变化和发展的过程，而散文的情绪变化往往与散文的主题紧密相连。而对于读者，可以通过理解作者的情绪变化，使自身更能全面、深刻地理解散文的主旨和作者的观点。

正如图1所示，文章开始描述了荷塘“已不是以往模样”，呈现了对时光流逝带来的感慨。然而，随着“平静地浮在水面的叶子已亭亭玉立于荷塘”，使荷塘展现出了美丽和生机，作者对它的喜爱由此而生。在文章的后半部分，作者揭示了周围人们的行为，特别是对年轻女子的描述“装扮不予年龄相仿，裸露着玉腿，半舒着胸房”，表达了作者对社会道德和价值观的不满和担忧，情绪逐渐转向厌恶。

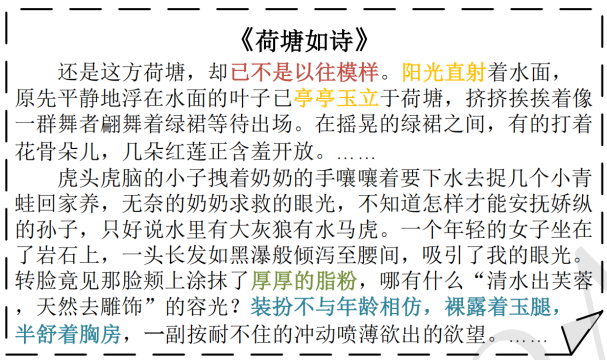



图 1: 示例《荷塘如诗》

《荷塘如诗》一文展示了散文的典型特点，而目前的方法存在不足之处。首先，散文常采用隐晦的语言和象征手法来表达情绪，比如“已不是以往模样”、“厚厚的脂粉”等，现有的情绪分析方法难以准确捕捉其中的隐含情绪(Zhang and Qian, 2020)。其次，散文中充满了丰富的意象，比如“荷塘”、“叶子”等，但现有方法往往缺乏对意象的深层表示，无法充分挖掘其中的情绪信息，从而影响了情绪分析结果的准确性。此外，散文通常具有复杂的组织结构，包含多个段落和句子，文章的长度可能达到30多个句子。然而，现有方法在理解文章整体思路和内在联系方面的能力有限(Tran and Phan, 2020)，无法全面把握文章的意义和情绪变化。

因此，为了提高对散文情绪的准确理解和深度分析能力，需要充分考虑散文特点对散文情绪变化进行分析。本文利用意象知识库，并通过注意力机制将意象引入到句子表示中，以丰富句子的语境信息。同时，构建上下文依赖图CotGAT和语篇依赖图DepGAT，对全局上下文及结构信息进行建模，获取全局的上下文理解。

本文的主要贡献如下：

(1) 构建了一个全新的散文情绪变化数据集，为本任务的研究者提供数据支持。

(2) 提出了基于双图注意力网络的篇章级散文情绪变化分析方法。通过意象知识库、上下文依赖图、语篇依赖图以及愉悦效价识别，获得丰富的句子表示，增强了模型的情绪理解能力。

(3) 在散文情绪变化数据集上，与对比方法的实验结果表明，本文方法具有较好的性能。

## 2 相关工作

散文情绪变化分析研究的基础是剖析文章的情绪，因此，情绪变化分析技术关联于文本情绪分类技术，涉及到词语级、句子级和篇章级文本情绪分析。Zhao等人(2022)提出情绪词选择这一情绪感知辅助任务，旨在让模型学习从给定的选项中选择正确的情绪词。Tian等人(2020)提出了SKEP模型，该模型随机地将输入文本中的一些方面项-情绪词对替换为MASK，并进行预测。而Huang等人(2019)采用滑动窗口的方法，为每个输入句子生成一个文本级图，其中，句子中的每个单词与其左右相邻的p个词相连。散文中的情绪常以含蓄的方

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版
* 通讯作者　Corresponding Author.
**基金项目**：国家自然科学基金项目(62076158,62106130,62376143)；山西省基础研究计划项目(20210302124084)；山西省高等学校科技创新项目(2021L284)

式表达，语境信息对于散文情绪的探究至关重要。然而，句子级情绪分析往往忽视了文章中句子间的复杂关系，限制了对文章的全面理解。

与句子级情绪分析不同，篇章级情绪分析考虑到文本的整体语境和结构，以及句子之间的关系。Yao等人(2019)为整个数据集构建了一个异构图，其利用词的共现信息建立边，在词节点与文档节点之间利用词的词频和词的文档词频建立边。Abreu等人(2019)将注意力机制应用于句子级上下文向量，对句子间的关系进行建模。Dai等人(2022)使用预训练的RoBERTa模型对每个句子进行编码，并且使用两层Transformer捕获不同句子之间的交互。这些方法都对上下文关系进行了建模，但缺乏对语篇依赖关系的考虑。然而，在对散文进行情绪分析时，需要充分考虑文章结构信息，才能精确地把握作者的情绪。因此，本文选择构建CotGAT和DepGAT，分别建模上下文关系和语篇依赖信息，以较全面地理解文章结构并把握文章情绪，从而提升散文情绪变化的分析性能。

## 3 散文情绪变化数据集构建

由于本文的任务是面向散文的情绪变化分析方法，任务较新，目前并没有公开的标注数据集，因此，我们构建了一个全新的散文情绪变化数据集。具体来说，首先搜集了290篇散文作为初始数据，来源于近几年全国各地的高考语文阅读理解题目和“中国散文网”。其中，“中国散文网”是以散文团体、散文作家、散文报刊及散文爱好者为主要服务对象，集文化资讯、散文报道、佳作交流、散文比赛、编辑翻译等于一体的大型文学网站。下面详细介绍数据集的构建过程。

### 3.1 标注体系的制定

为了建立散文情绪变化分析，首先制定了一个标注体系，旨在确保数据集的规范性、通用性和可扩展性。本任务涉及五种标注信息：(1)愉悦效价类别(PVC)；(2)情绪类别(SC)；(3)情绪(S)；(4)情绪变化(SCH)；(5)情绪变化的原因(RSCH)。

标注者按照以上五种标注信息的顺序进行标注。对于每一条数据，ID代表“文章编号_句子编号”，例如，“2_0”代表本条数据为文章2的第一个句子。Russell和Mehrabian的PAD情绪三维理论中指出，效价又称愉悦度(Pleasure-displeasure)，用于表示情绪状态的正负性，是衡量情绪的重要维度。对于情绪类别(SC)，我们借鉴Demszky等人(2020)的方法，结合《礼记》中的“七情”以及《白虎通》中的“六情”，综合散文的情绪特点，共分为十三类，包括乐、爱、敬、赞、愿、怒、哀、惧、恶、惊、信、羞、其他。因此，对于愉悦效价类别(PVC)，参照PAD三维理论和圆环模型(Russell, 1980)，通过情绪极性划分为正面、中性、负面，随后将十三类的情绪映射到这三类，形成了一个层次结构，具体结构如图2所示。在这个结构中，前三类提供了情绪极性价，有助于识别句子中存在的情绪。例如，“况且，湖景若无可观，湖中的出产可是很名贵呀”这个句子唤起了正面的情绪效价，有助于从底部的六个类别中识别情绪。

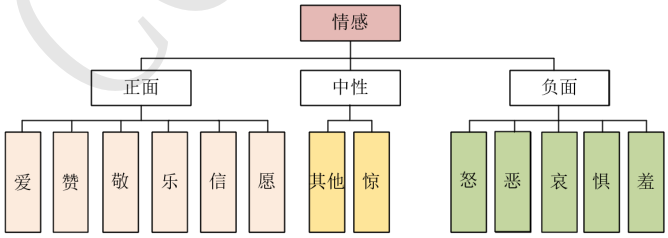



图 2: 情绪层次结构

在标注情绪类别(SC)的同时，我们也对文章中细节情绪(S)进行了标注，例如“喜爱”、“厌恶”等，有助于加深对文章情绪的理解。情绪变化(SCH)中的“YES”表示文章在该文本处发生情绪变化，而“NO”则表示没有发生变化；情绪变化的原因(RSCH)则记录了情绪变化(即SCH=“YES”)的详细原因。

### 3.2 数据标注

为了保证数据的准确性和可用性，本文首先对收集到的文章进行句子级的划分，以保证每条数据包含较为完整的语义信息；其次，由人工对句子进行合并或删除，以进一步提高数据的

质量；最后，过滤数据中的特殊符号、网页地址等冗余内容，以保证数据的准确性和可用性。至此，获得散文情绪变化分析的初始数据。在此基础上，借助ChatGPT[0]进行“预标注”。

ChatGPT是一个基于GPT3.5的自然语言处理模型，具有出色的语言理解和生成能力，可以应用于多种自然语言处理任务，如情绪分析、语言理解等。为了提高标注效率及准确率，由ChatGPT和两位具有良好的文学素养和语言学背景的标注者共同标注。为了使用ChatGPT，构建两种提示如下：

（1）请按照情绪类别，依次输出每个句子的情绪类别及具体的情绪，比如乐-快乐。注意只需要列出最主要的情绪类别。其中，情绪类别分为13类，分别为：乐，爱，敬，赞，愿，怒，哀，惧，恶，惊，信，羞，其他。文章：CONTENT。

（2）给出文章CONTENT的情绪脉络及变化的原因。

利用上述两类提示，借助ChatGPT生成相应的标注结果，在此基础上，针对每一条数据，由两位标注者分别进行“再标注”，对于标注不一致的样例，通过多数投票机制获得最终的标注结果。经过数据标注，总共整理出7788条散文情绪变化分析数据，部分样例如下表1所示。

| ID | Text | PVC | SC | S | SCH | RSCH |
|---|---|---|---|---|---|---|
| 2_0 | 还是这方荷塘，却已不是以往模样。 | 负面 | 哀 | 感慨 | NO | NONE |
| 2_1 | 阳光直射着水面，原先平静地浮在水面的叶子已亭亭玉立于荷塘，挤挤挨挨着像一群舞者翩舞着绿裙等待出场。 | 正面 | 爱 | 喜爱 | YES | 描述阳光直射和叶子亭亭玉立，显示了荷塘的美丽和生命力的提升，体现了喜爱的情绪。 |
| 2_2 | 在摇晃的绿裙之间，有的打着花骨朵儿，几朵红莲正含羞开放。 | 正面 | 爱 | 喜爱 | NO | NONE |

表 1: 散文情绪变化数据集标注样例

通过对以上标注结果的统计，两组标注者的标注结果的Kappa值为85.54%，表现了标注结果的一致性和可靠性；人工标注的情绪类别与GPT标注的情绪类别的Kappa值为32.71%，说明ChatGPT在理解文章情绪具有局限性。

## 4 双图注意力网络的情绪变化分析模型

针对散文情绪变化分析任务，本文综合意象知识库，提出了基于双图注意力网络的篇章级散文情绪变化分析模型。整个模型包括五个主要部分：特征提取层、愉悦效价识别层、DualGATs层、情绪预测层和情绪变化识别层。首先，利用意象知识库，获得基于意象知识增强的句子表示；进一步，利用上下文依赖图、语篇依赖图以及相互交叉注意机制进行多层地提取和迭代，获得基于上下文知识增强的句子表示、基于语篇依赖信息增强的句子表示。其次，将这些表示添加愉悦效价识别层，通过线性变化和ReLU函数，获得基于愉悦效价知识增强的句子表示。最后，将以上三者进行拼接，通过全连接网络对各个句子的情绪类别进行预测，得到最终的情绪变化结果。模型的整体框架如图3所示。其中，散文$D=[S_1,S_2,...,S_N]$，$N$表示散文中句子的数量，$S_i$为第$i$个句子，$y_i \in Y$和$y_i^v \in V$分别是情绪标签和愉悦效价，$Y$和$V$分别为情绪标签集合和愉悦效价集合。

### 4.1 特征提取层

本模型使用预训练RoBERTa-large[1]模型，将散文$D=[S_1,S_2,...,S_N]$(图3中$N$=5)中的每个句子$S_i$进行编码，获得$N$个句子的嵌入表示分别为$h_1^s,h_2^s,...,h_N^s$。由于意象知识对句子的情绪识别具有重要的指导作用，因此，利用课题组建立的意象知识库$KB=(k_1,k_2,...,k_M)$，查询句子中是否存在意象，对存在意象的句子，进行再次嵌入表示$h_i^k$。最后，使用注意力机制，为两种表示分配权重，融合句子及意象知识的特征表示为$h_i$。该过程如图4所示，具体操作如下：

[0] https://chat.openai.com

[1] https://huggingface.co/hfl/chinese-roberta-wwm-ext-large


(c) 2024 中国中文信息学会计算语言学专业委员会

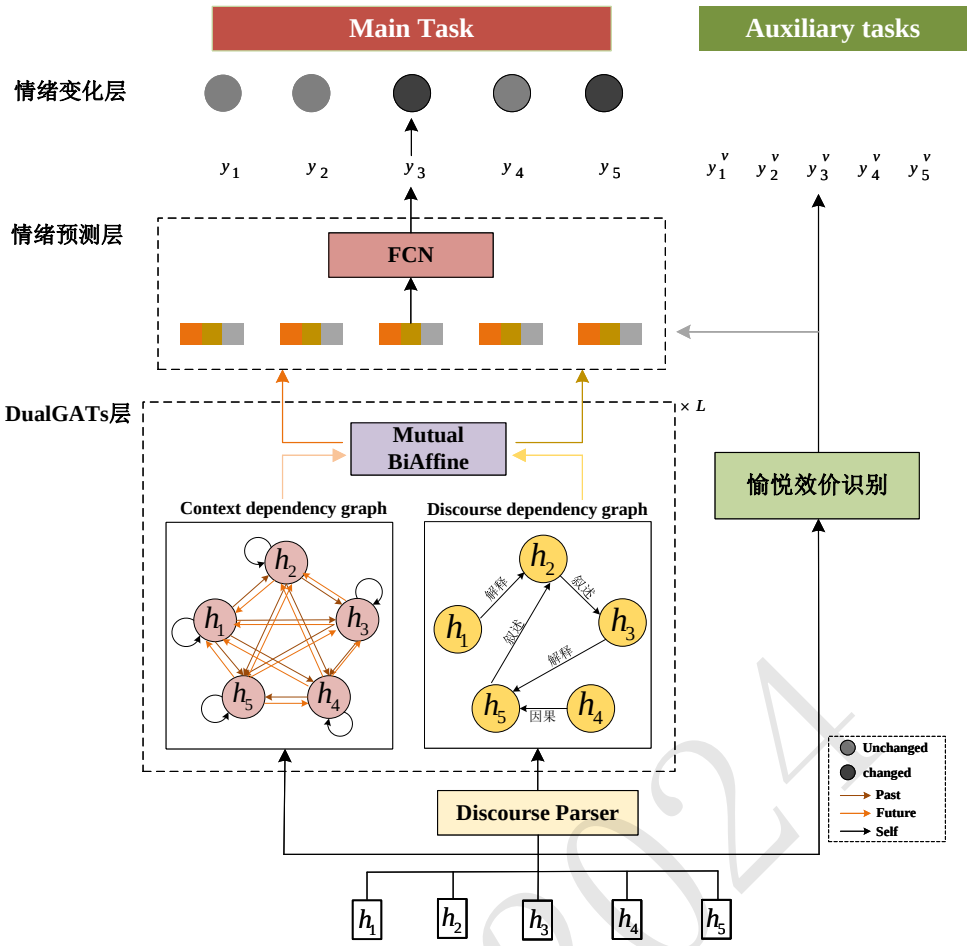



图 3: 散文情绪变化分析模型的框架图

$$a = V^T \tanh(W[h_i^s || h_i^k]) \tag{1}$$

$$h_i = \alpha \begin{bmatrix} h_i^{sT} \\ h_i^{kT} \end{bmatrix} \tag{2}$$

其中，$h_i \in R^{d_s}$，$d_s$表示特征表示的维度，所有句子的特征表示为$H^s \in R^{N \times d_s}$。

## 4.2 愉悦效价识别层

依据图2中提出的情绪层次结构，本文设计愉悦效价识别层，利用句子的特征表示$h_i$作为输入，通过公式(3)，获得给定句子的愉悦分布，以期使用该分布提升主要的情绪变化分析任务的性能。

$$val_i^v = \mathrm{ReLU}(W_h^v h_i + b_h^v) \tag{3}$$

其中，$W_h^v \in R^{d_h \times d_s}$，$b_h^v \in R^{d_h}$是可训练的参数，$val_i^v \in R^{d_h}$表示句子$S_i$的愉悦效价分布。通过softmax函数，对散文$D$中每个句子$S_i$的愉悦效价类别进行分类：

$$\hat{p}_i^v = \mathrm{softmax}(W_l^v val_i^v + b_l^v) \tag{4}$$

$$\hat{y}_i^v = \arg\max_i(\hat{p}_i^v) \tag{5}$$

其中，$W_l^v \in R^{d_e \times d_h}$，$b_l^v \in R^{d_e}$是可训练的参数，$\hat{y}_i^v$表示预测的句子$S_i$的愉悦效价标签。


(c) 2024 中国中文信息学会计算语言学专业委员会

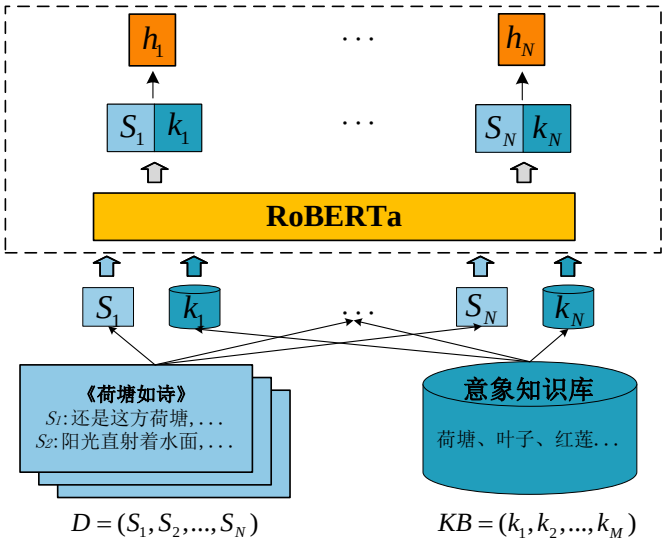



图 4: 特征提取层的概述图

### 4.3 DualGATs层

为了有效地捕获文章中的上下文知识和语篇依赖结构，本文设计了DualGATs层，包括三个基本模块：上下文依赖图、语篇依赖图及交互模块。本节概述初始层中每个模块的计算过程，随后将其扩展到多个后续层。

(1) 上下文依赖图

通过构建上下文依赖图(CotGAT)，能够较好地理解文章的情绪变化过程。CotGAT的构建基于三种上下文依赖类型：Past、Future和Self（简称为集合$R^{Cot}$）。其中，Past表示过去句子对当前句子的影响，有助于理解前文情绪对当前情绪的延续和转折；Future指的是将来句子对当前句子的影响，反映后续内容对当前情绪的预设和引导作用；Self表示当前句子内部的情绪变化和发展。

本文将文章的上下文依赖图定义为$G^{Cot} = (V^{Cot}, E^{Cot})$，其中$V^{Cot}$表示节点集合，$E^{Cot}$是描述节点(句子)之间的上下文依赖关系的邻接矩阵。$v_i^{Cot}$表示$S_i$(第i个句子)，其表示用相应的特征表示$h_i$进行初始化。对于任何$S_i$和$S_j$，如果它们满足上下文依赖类型$r^{Cot} \in R^{Cot}$，那么$e_{i,j}^{Cot} = r^{Cot}$。构建上下文依赖图之后，应用CotGAT模块在图节点间传播和聚合上下文信息。CotGAT模块建立在GAT(Velickovic et al., 2018)之上，但包含类型编码，以考虑节点之间的时间依赖类型。具体来说，对于给定节点$v_i^{Cot}$，CotGAT对其相邻节点进行信息聚合，具体操作如式(6)-(7)所示。

$$a_{i,j} = \text{softmax}(\text{LRL}(a^T[Wh_i||Wh_j||e_{i,j}^{Cot}])) \tag{6}$$

$$h_i^{Cot} = \sum_{j \in N_i^{Cot}} a_{i,j} W h_j \tag{7}$$

$a_{i,j}$表示从节点$v_i^{Cot}$到邻居$v_j^{Cot}$的边权值，LRL表示LeakyReLU激活函数，$W$和$a$表示可训练的参数，$e_{i,j}^{Cot} \in R^{|R^{Cot}|}$表示节点$v_i^{Cot}$和$v_j^{Cot}$之间的语篇依赖类型对应的one-hot编码（模型训练时固定），$|R^{Cot}|$表示语篇依赖类型的数量，$||$表示连接操作，$N_i^{Cot}$表示图$G^{Cot}$中节点$v_i^{Cot}$的邻居，$h_i^{Cot} \in R^{d_h}$表示节点$v_i^{Cot}$在CotGAT模块更新后的隐藏表示。更新后的所有节点的隐藏表示可以表示为$H^{Cot} \in R^{N \times d_h}$。将初始层的CotGAT计算过程进行总结，如式(8)所示。

$$H^{Cot} = \text{CotGAT}(H^s, E^{Cot}) \tag{8}$$

(2) 语篇依赖图

文章的语篇结构反映了内容的逻辑框架，也是情绪变化分析的重要依据。根据Asher等人(2016)中的概述，依赖逻辑类型共十五种：注释、澄清问题、解释、背景、叙述、确认、细

化、延续、条件、问答对、交替、结果、纠正、平行和对比（简称为集合$R^{Dep}$）。这些依赖类型在文章中扮演着不同的角色。注释、澄清问题、解释、背景、叙述和确认这六种类型，提供了情绪变化的背景信息和解释说明，有助于更好地理解情绪变化的背景和原因。细化和延续这两种类型延续了情绪变化的主题，有助于深入探讨情绪变化的细节。此外，条件和问答对的运用，促进了对于情绪变化的思考和探索。而交替、结果、纠正、平行和对比这五种类型，则呈现不同情绪状态之间的对比和转变。因此，本文利用这些依赖逻辑类型，建立语篇依赖图(DepGAT)，以刻画情绪变化的内在逻辑。

在本方法中，文章中的每个句子都被视为一个基本语篇单元(记作EDUs)。因此，本文将文章的语篇依赖图定义为$G^{Dep} = (V^{Dep}, E^{Dep})$，其中$V^{Dep}$表示EDUs的节点集，$E^{Dep}$表示EDUs之间语篇依赖关系的邻接矩阵。$v_i^{Dep}$表示与句子$S_i$对应的节点，其表示用相应的特征表示$h_i$进行初始化。使用预训练的解析器(Shi and Huang, 2019)预测文章中各个句子之间的语篇依赖关系。如果从句子$S_i$到句子$S_j$，存在语篇依赖类型$r^{Dep} \in R^{Dep}$，边$E_{i,j}^{Dep}$被赋予相应的依赖类型。构建文章的语篇依赖图后，应用DepGAT模块在图节点之间传播和聚合句子依赖信息。同样，总结初始层DepGAT的计算过程，如式(9)所示。

$$H^{Dep} = \text{DepGAT}(H^s, E^{Dep}) \tag{9}$$

(3) 交互模块

为了有效融合上下文知识和语篇依赖信息，采用相互交叉注意机制，建立融合语篇依赖结构的上下文表示、融合上下文信息的语篇结构表示。具体计算见式(10)-(12)所示。

$$A_1 = \text{softmax}(H^{Cot} W_1 (H^{Dep})^T) \tag{10}$$

$$A_2 = \text{softmax}(H^{Dep} W_2 (H^{Cot})^T) \tag{11}$$

$$H^{Cot'}, H^{Dep'} = A_1 H^{Dep}, A_2 H^{Cot} \tag{12}$$

其中，$W_1, W_2 \in R^{d_h \times d_h}$是可学习的参数，$A_1, A_2 \in R^{N \times N}$分别是从$H^{Cot}$到$H^{Dep}$，从$H^{Dep}$到$H^{Cot}$的变化矩阵，$H^{Cot'}, H^{Dep'} \in R^{N \times d_h}$分别是所有句子的融合语篇依赖结构的上下文表示和融合上下文信息的语篇结构表示。

(4) DualGATs层的多次迭代

为了深度提取句子的上下文知识和语篇依赖信息，使用$L$层的DualGATs进行迭代。具体过程如式(13)-(16)所示。

$$H^{Cot,[l]} = \text{CotGAT}(D^{[l]}, E^{Cot}) \tag{13}$$

$$H^{Dep,[l]} = \text{DepGAT}(S^{[l]}, E^{Dep}) \tag{14}$$

$$H^{Cot',[l]}, H^{Dep',[l]} = Inter(H^{Cot,[l]}, H^{Dep,[l]}) \tag{15}$$

$$D^{[l+1]}, S^{[l+1]} = H^{Cot',[l]}, H^{Dep',[l]} \tag{16}$$

其中，$D^{[0]} = S^{[0]} = H^s$，$l \in [0, L-1]$。

### 4.4 情绪预测层

通过连接第$L$层DualGATs的输出和愉悦效价分布得到$S_i$的最终表示，通过全连接网络进行分类，具体操作如式(17)-(19)所示。

$$l_i = \text{ReLU}(W_h[h_i^{Cot',[L]} || h_i^{Dep',[L]} || val_i^v] + b_h) \tag{17}$$

$$\hat{p}_i = \text{softmax}(W_l l_i + b_l) \tag{18}$$

$$\hat{y}_i = \arg\max_i(\hat{p}_i) \tag{19}$$

其中$\hat{y}_i$是预测的句子$S_i$的情绪标签，$h_i^{Cot',[L]}, h_i^{Dep',[L]} \in R^{d_h}$表示$H^{Cot',[L]}$和$H^{Dep',[L]}$中的第$i$个表示，$val_i^v \in R^{d_h}$表示句子$S_i$的愉悦效价分布，$W_h \in R^{d_h \times 3d_h}$，$W_l \in R^{d_e \times d_h}$，$b_h \in R^{d_h}$，$b_l \in R^{d_e}$是全连接网络的可学习参数。


(c) 2024 中国中文信息学会计算语言学专业委员会

对于情绪标签识别与愉悦效价识别两个任务，采用交叉熵损失进行约束，损失函数分别如式(20)、(21)所示。

$$L_e = -\frac{1}{BN}\sum_{\beta=1}^{B}\sum_{i=1}^{N}(y_{\beta,i}\log\hat{p}_{\beta,i} + (1-y_{\beta,i})\log(1-\hat{p}_{\beta,i})) \tag{20}$$

$$L_v = -\frac{1}{BN}\sum_{\beta=1}^{B}\sum_{i=1}^{N}(y^v_{\beta,i}\log\hat{p}^v_{\beta,i} + (1-y^v_{\beta,i})\log(1-\hat{p}^v_{\beta,i})) \tag{21}$$

其中，$B$是文章的数量，$N$是第$\beta$篇文章的句子总数，$\hat{p}_{\beta,i}$，$\hat{p}^v_{\beta,i}$为预测的句子$S_i$的愉悦效价及情绪的标签分布。

利用多任务学习的方式，联合情绪标签识别和愉悦效价识别两者的损失得到最终损失$L$:

$$L = L_e + L_v \tag{22}$$

### 4.5 情绪变化识别层

在获得文章$D$中各个句子的情绪标签$y_1, y_2, ..., y_N$之后，对文章的情绪脉络具有清晰的认知。通过比较相邻句子的情绪类别，可以揭示情绪变化的位置和趋势。正如图5所示，随着文章的叙述，《荷塘如诗》经历了从感慨到喜爱，再到厌恶的情绪变化。

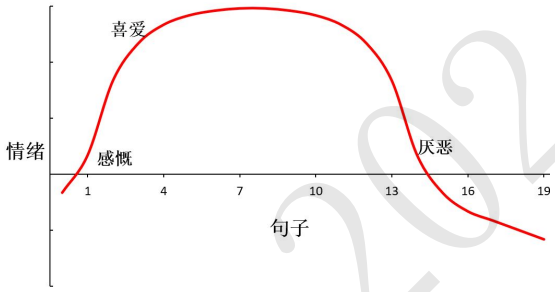



图 5: 篇章情绪变化分析图

## 5 实验

### 5.1 数据集介绍

根据人工标注整理后的7788条散文数据，按照7:1:2的比例进行随机划分，分为训练集、验证集和测试集。表2展示了数据集的具体划分情况。

| 划分 | 训练 | 验证 | 测试 | 总计 |
|---|---|---|---|---|
| 文章 | 200 | 30 | 60 | 290 |
| 句子 | 5203 | 1743 | 842 | 7788 |
| 变化 | 946 | 282 | 145 | 1373 |
| 文章长度(平均) | 26.02 | 28.06 | 29.05 | 26.85 |

表 2: 数据集划分介绍

### 5.2 评价指标及参数设置

为了多维度评价模型的性能，本文从情绪类别和愉悦效价两个方面，采用准确率（Accuracy）以及Macro-F1作为主要评价指标。本方法使用Adam优化器进行训练。使用验证集的F1得分对DaulGAT层数进行超参数搜索，范围为{1, 2, 3, 4, 5, 6}。对于其余超参数，学习率为5e-05，Drouput率为0.4，RoBERTa的特征表示的维数为1024，隐藏表示的维数为300，最大输入文本长度设置为128。使用验证集中的最佳检查点来评估测试集。所有的实验都是在一个具有40GB内存的NVIDIA A100 GPU上进行的，并且平均超过五次运行。

CCL 2024

### 5.3 实验设计

鉴于本文提出的任务较为新颖，目前还没有针对该任务的相关方法。同时，由于篇章级情绪分类方法相对较少，为验证本文提出方法的有效性，我们主要采用近年来用于句子级情绪分类任务的方法作为对比。在散文情绪变化数据集上将本文的方法与以下八种方法进行比较。

RoBERTa-large(Liu et al., 2019)：该模型对整个词进行MASK处理，模型能够充分理解上下文及语义关联。

ERNIE[2]：该模型基于BERT(Devlin et al., 2019)模型，在预训练阶段增加外部知识，实现了实体级别的掩码机制。

Text-level GNN(Huang et al., 2019)：该模型没有构建单个语料库级图，而是为每个输入文本生成一个文本级图，节点从相邻节点获取信息以更新其表示。

Vanilla-VGCN-BERT：类似于(Jeong et al., 2019)，通过BERT和VGCN产生两个独立的表示，并将它们连接起来。最后对连接表示应用ReLU和全连接层以进行分类。

VGCN-BERT(Lu et al., 2020)：该模型通过注意机制实现局部信息与全局信息的交互，允许它们相互影响，并共同构建用于分类的最终表示。

SKEP(Tian et al., 2020)：该模型构建了三个情绪知识预测目标，将词、极性和方面层面的情绪信息嵌入到预训练的情绪表示中。

SentiWSP(Fan et al., 2022)：该模型通过设计对比学习框架，提升已训练的判别器对于句子情绪层面的信息捕捉能力。

KESA(Zhao et al., 2022)：该模型提出情绪词选择和条件情绪预测两个情绪感知辅助任务，并相应地将它们整合到下游任务的目标中。

### 5.4 实验结果及分析

#### 5.4.1 实验比较及分析

利用5.3节设计的八种方法与本文方法进行对比实验，实验结果如表3所示。

| 模型 | 情绪(%) | | 愉悦效价(%) | |
|---|---|---|---|---|
| | Acc | MacF1 | Acc | MacF1 |
| RoBERTa-large | 48.22 | 45.78 | 62.31 | 63.14 |
| ERNIE | 37.89 | 32.30 | 64.13 | 60.77 |
| Text-level GNN | 18.81 | 18.47 | 44.62 | 44.01 |
| Vanilla-VGCN-BERT* | 48.40 | 45.47 | 62.42 | 62.55 |
| VGCN-BERT* | 48.64 | 45.56 | 62.77 | 62.67 |
| SKEP* | 48.41 | 44.15 | 63.30 | 63.08 |
| SentiWSP | 47.82 | 46.10 | 60.19 | 58.45 |
| KESA | 50.45 | 45.65 | 66.43 | 61.63 |
| Ours | **52.61** | **51.60** | **77.57** | **77.27** |

表 3: 情绪及愉悦效价识别结果(标记*表示使用RoBERTa-large模型)

通过表3可以看出：

(1) 本文提出的模型在散文数据集上始终优于所有对比方法。这些结果表明，本文的方法可以有效分析散文情绪。

(2) SKEP模型使用情绪词、方面项进行预训练，其性能显著优于ERNIE等没有使用这些知识的方法。因此，本文利用意象知识库，获取意象信息，从而较为准确地了解情绪来源。

(3) 本文提出的方法明显优于SentiWSP、VGCN-BERT等句子级情绪分析方法，其原因在于本文方法依据散文的复杂性和逻辑性，充分考虑句间信息，对上下文关系及语篇依赖关系进行建模，从而提升了模型的效果。

#### 5.4.2 消融实验

为了验证模型中每个组件的有效性，我们进行了广泛的消融研究。将模型去除上下文依赖图记为-CotGAT；将模型去除语篇依赖图记为-DepGAT；将模型去除图注意力网络间的交互模块记为-交互模块；将模型去除意象知识库记为-意象；将模型去除愉悦效价识别模块记为-愉悦效价识别。结果如表4所示。

[2]https://github.com/PaddlePaddle/ERNIE


(c) 2024 中国中文信息学会计算语言学专业委员会

| 模型 | 情绪(%) | | 愉悦效价(%) | |
|---|---|---|---|---|
| | Acc | MacF1 | Acc | MacF1 |
| Ours | **52.61** | **51.60** | **77.57** | **77.27** |
| -CotGAT | 51.12 | 50.84 | 76.25 | 76.00 |
| -DepGAT | 50.66 | 51.01 | 76.42 | 76.45 |
| -交互模块 | 51.41 | 50.70 | 75.62 | 75.38 |
| -意象 | 50.37 | 50.09 | 77.40 | 77.02 |
| -愉悦效价识别 | 51.04 | 50.43 | 76.49 | 76.47 |

表 4: 消融实验结果

由表4可以看出：

（1）本文的方法（Ours）是所有性能中最佳的。从完整的模型中删除交叉注意模块、意象知识库、愉悦效价识别层的任何一个，性能都会下降。其中意象模块对散文数据集的影响最大，可能是由于散文是以意象为线索，将情绪贯穿于文章中。

（2）-DepGAT仅对文章的上下文关系进行建模，而散文具有含蓄细腻的特点，它获取的句子表示比较浅层。

（3）-CotGAT只对文章的语篇依赖关系进行建模，并且取得了比DepGAT更高的准确率，说明语篇依赖结构有助于提取深层句间信息。

### 5.4.3 DualGAT层数的影响

为了研究DualGAT层数的影响，我们在散文情绪变化数据集上对一到六层的DualGAT进行评估。如图6所示，具有两个DualGAT层的模型表现最佳。一方面，在层数较少的情况下，上下文信息和语篇依赖信息可能无法得到充分的提取。另一方面，当层数过多时，产生了冗余或者兼容的信息，导致模型的性能下降。

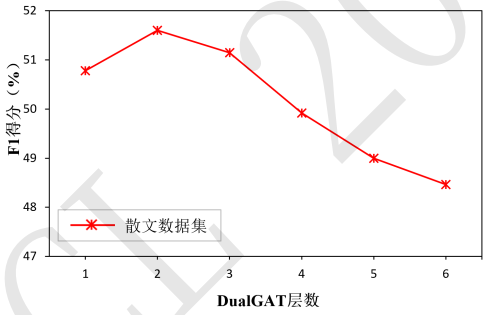



图 6: DualGAT层数的影响

### 5.4.4 错误分析

在本节对错误结果进行分析。表5列出一些本文模型未能正确识别出情绪类别的例子。

| 文章号 | 例句 | 情绪类别 | 模型识别 |
|---|---|---|---|
| 12 | 而突然冒出的尖厉噪声当然属于鸣蝉，这些心里只有自己的家伙个个愤世嫉俗、怀才不遇，不停歇地嚷着:“是我——是我——”。 | 恶 | 怒 |
| 225 | 实际，我向来对于花木无所爱好。 | 其他 | 恶 |

表 5: 部分情绪类别识别错误示例

根据表5，在文章编号为12的一个句子中，“恶”与“怒”都属于负面情绪，但两者代表着不同的含义，而模型无法界定两者的区别，导致结果有误。在文章编号为225的一个句子中，模型错误地将情绪识别为“恶”，可能是由于“向来”、“无所爱好”等具有负面情绪色彩的词语，然而结合上下文信息，该句并没有表达明确的情绪倾向。


(c) 2024 中国中文信息学会计算语言学专业委员会

## 6 结论

本文利用意象知识库，提出一个基于双图注意力网络的篇章级散文情绪变化分析方法，同时设计愉悦效价识别层，可以较精确地对散文情绪变化进行分析。其中为了有效地交互上下文依赖图与语篇依赖图之间的信息，使用相互交叉注意机制。在散文数据集上的大量实验表明，本模型可以有效地建模文章信息，提升对情绪变化分析的能力，为散文阅读理解中的思想情绪变化类问题的解答提供支持。

## 参考文献

Jader Abreu, Luis Fred, David Macêdo, and Cleber Zanchettin. 2019. Hierarchical attentional hybrid neural networks for document classification. In *Artificial Neural Networks and Machine Learning – ICANN 2019: Workshop and Special Sessions: 28th International Conference on Artificial Neural Networks*, page 396–402.

Nicholas Asher, Julie Hunter, Mathieu Morey, Benamara Farah, and Stergos Afantenos. 2016. Discourse structure and dialogue acts in multiparty dialogue: the STAC corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 2721–2727.

Xiang Dai, Ilias Chalkidis, Sune Darkner, and Desmond Elliott. 2022. Revisiting transformer-based models for long document classification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7212–7230.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Shuai Fan, Chen Lin, Haonan Li, Zhenghao Lin, Jinsong Su, Hang Zhang, Yeyun Gong, JIan Guo, and Nan Duan. 2022. Sentiment-aware word and sentence level pre-training for sentiment analysis. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4984–4994.

Lianzhe Huang, Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2019. Text level graph neural network for text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3444–3450.

Chanwoo Jeong, Sion Jang, Hyuna Shin, Eunjeong Lucy Park, and Sungchul Choi. 2019. A context-aware citation recommendation model with bert and graph convolutional networks. *Scientometrics*, 124:1907–1922.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Zhibin Lu, Pan Du, and Jianyun Nie. 2020. Vgcn-bert: Augmenting bert with graph embedding for text classification. *Advances in Information Retrieval*, 12035:369–382.

James A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178.

Zhouxing Shi and Minlie Huang. 2019. A deep sequential model for discourse parsing on multi-party dialogues. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7007–7014.

Hao Tian, Can Gao, Xinyan Xiao, Hao Liu, Bolei He, Hua Wu, Haifeng Wang, and Feng Wu. 2020. SKEP: Sentiment knowledge enhanced pre-training for sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4067–4076.

Thien Khai Tran and Tuoi Thi Phan. 2020. Capturing contextual factors in sentiment classification: An ensemble approach. *IEEE Access*, 8:116856–116865.


(c) 2024 中国中文信息学会计算语言学专业委员会

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio', and Yoshua Bengio. 2018. Graph attention networks. *International Conference on Learning Representations*.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Graph convolutional networks for text classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):7370–7377.

Mi Zhang and Tieyun Qian. 2020. Convolution over hierarchical syntactic and lexical graphs for aspect level sentiment analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 3540–3549.

Qinghua Zhao, Shuai Ma, and Shuo Ren. 2022. KESA: A knowledge enhanced approach to sentiment analysis. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 766–776.

王小环. 2023. 论中国现代散文语言的美学呈现. 青岛科技大学学报:社会科学版, 39(01):59–65.

CCL 2024

# 基于主题模型与图神经网络的突发公共卫生事件国际舆情演化分析研究

**高境健[1]，桑国明[1]，刘智[1]，张益嘉*[,1]，林鸿飞[2]**
1.大连海事大学，信息科学技术学院，辽宁大连，116000
2.大连理工大学，计算机科学与技术学院，辽宁大连，116000
gjj@dlmu.edu.cn,sangguoming@dlmu.edu.cn
liuzhi@dlmu.edu.cn,zhangyijia@dlmu.edu.cn,hflin@dlut.edu.cn

## 摘要

研究突发公共卫生事件国际舆情演变规律，对国际舆情资源进行应急管理和舆论疏导有重要借鉴价值。本文使用谷歌新闻数据库以各国针对COVID-19的报道为对象，构建国际舆情数据集。采用主题模型、图神经网络模型，结合时间、空间维度与舆情生命周期探究全球舆论主题-情感的演化态势，模型准确率为0.7973，F1值为0.7826，性能优于其他基线模型。研究发现，各国舆情呈现放射传播状态。国际媒体舆论的情感倾向和讨论主题存在正相关且随时间进行转变。

**关键词：** 舆情分析；突发公共卫生事件；主题模型；图神经网络；时空演化分析

# International Public Opinion Evolution Analysis on Sudden Public Health Events using Topic Model and Graph Neural Network

**Jingjian Gao[1], Guoming Sang[1], Zhi Liu[1], Yijia Zhang*[,1], Hongfei Lin[2]**
1.Dalian Maritime University ,Dalian, Liaoning 116000, China
2. Dalian University of Technology, Dalian, Liaoning 116000 ,China
gjj@dlmu.edu.cn,sangguoming@dlmu.edu.cn
liuzhi@dlmu.edu.cn,zhangyijia@dlmu.edu.cn,hflin@dlut.edu.cn

## Abstract

Studying the evolution of international public opinion in sudden public health emergencies has important reference value for emergency management and public opinion guidance of international public opinion resources. This article uses Google News database to construct a international public opinion dataset based on reports from various countries on COVID-19. Using topic model and graph neural network model, combined with time and space dimensions and public opinion lifecycle, the evolution trend of global public opinion topic emotion was explored. The accuracy of the model was 0.7973, with an F1 value of 0.7826, which outperformed other baseline models. Research has found that public opinion in various countries is in a radiating state of spread. The emotional tendencies and discussion topics of international media public opinion are positively correlated and change over time.

**Keywords:** Public opinion analysis , Sudden public health emergencies , Topic model , Graph neural network , Spatiotemporal evolution analysis

## 1 引言

*张益嘉（通讯作者）：zhangyijia@dlmu.edu.cn
基金项目：辽宁省社会科学规划基金项目“新冠病毒疫情下的国际舆情分析研究”（L20BTQ008）
©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

舆情是指群体针对特定突发事件所表达的态度、意见和情绪的总和。面对复杂的国际形势，各国针对公共卫生事件和社会保障事件等突发事件的态度有所不同，研究国际舆情有助于了解国际舆论风向，针对性地调整国家战略以应对可能的风险和危机。突发公共卫生事件具有突发性、急剧性、不可预测性等特征，往往会对人民群众的生命健康造成严重威胁，并引发公众恐慌、不满、愤怒等复杂情感(An et al., 2018)。媒体报道往往会带动公众情绪，从报道的情绪中识别公众意见倾向并融合主题信息，判断舆情的大体走向，把握舆情演化规律才能有效引导舆情(曾子明and 陈思语, 2023)。

自2020年初全球大范围的突发公共卫生事件“新冠肺炎疫情”的爆发以来，各国政府、媒体及大众对疫情相关信息高度关注。突发的国际舆情事件在公众关注度、情感态度、舆论导向等方面随时间动态演变，不同地区用户情感受该事件影响程度也有较大差异，并存在明显的跨地域传播。这都使得单一围绕时间-地点的信息切片，或仅面向单一主题的舆情分析无法完整反映突发公共卫生事件背景下的国际舆情演化全貌(边晓慧and 徐童, 2022)。

高效、全面、准确地分析国际舆情演化态势，对于政府在突发公共事件下有效疏导公共舆论具有重要价值。本文以新冠肺炎的国际媒体报道为研究对象，构建了国际舆情数据集，基于主题模型、图神经网络及舆情周期理论提出了国际舆情演化分析框架。采用多主题视角和全局视角分析舆情时空演化态势，挖掘舆论主题及情感倾向的跨地域动态扩散特征，进而使各国政府把握全球疫情焦点舆论和发展方向，结合本国舆情态势制定相关干预和引导舆论方向战略。

## 2 相关工作

### 2.1 舆情演化分析

舆情事件的演化往往具有生命周期(史波, 2010)。国内外研究者均对舆情的传播过程进行了探索，这些研究根据不同角度将舆情按照事件发生序列和发展生命周期来划分阶段并构建模型(甘宇祥et al., 2021)。Burkholder and Toole (1995)提出三阶段模型：潜伏期、扩散期、消退期。Canton (2006)基于三阶段模型结合医学知识提出著名的四阶段F模型：前驱期(Prodromal)、发作期(Breakout)、延续期(Chronic)、痊愈期(Resolution)。潘崇霞 (2011)提出简化的网络舆情演化三阶段模型：初始传播、迅速扩散、消退。上述采用生命周期理论对公共事件进行阶段划分的研究，对于挖掘网络舆情演化规律具有重要意义。

### 2.2 文本挖掘

随着文本挖掘技术的不断成熟，使用高效的文本挖掘技术处理大量的文本数据成为了主流，弥补了文本数据在计量分析和内容分析中需要大量人工的缺点。Huang et al. (2012)将LDA (Latent Dirichlet Allocation) 模型应用到微博短文本主题检测。Cheng et al. (2021)基于LDA模型对国务政策进行主题提取，研究政策主题演化。当前，LDA模型已经广泛应用于主题挖掘研究中，但该模型存在无法融合上下文信息的问题，导致主题挖掘性能欠缺。

### 2.3 情感分析

情感分析方法是国际舆情分析研究的关键技术，目前主要基于三种方法，即情感词典、机器学习和深度学习(史伟et al., 2022)。情感词典作为早期重要的情感分类工具，词典的精细程度决定了对文本情绪分类的准确性。曾子明and 万品玉 (2018)基于情感词典提出一种多层次融合模型，对文本进行正负情感的分类。使用情感词典进行情感分类有着明显的领域针对性以及泛化性较低的缺点。机器学习作为使用大量语料进行模型训练的方法，在情感分类领域有着较好的效果。邓君et al. (2020)基于Word2Vec和SVM对舆论进行二分类任务，该方法并没有利用语法结构的信息。Machuca et al. (2021)NLP技术融合回归算法针对舆情背景下的Twitter舆论进行分类。机器学习方法未考虑上下文语义信息对情感分类效果的影响。深度学习有着更好的特征提取能力，通过融合文本的上下文信息增强情感分类准确性。传统的深度学习模型在长期依赖问题中依旧乏力，Vaswani et al. (2017)构建Transformer模型，使用自注意力机制进行词向量嵌入对长期依赖问题有着较好的效果。为充分利用左右两侧的上下文信息，Devlin et al. (2018)基于微调的多层双向Transformer模型编码器构建了BERT(Bidirectional Encoder Representation from Transformers) 预训练模型，并用于提取文本特征，从而达到融合两侧信息的效果(Burkholder and Toole, 1995)。


(c) 2024 中国中文信息学会计算语言学专业委员会

综上，一方面，现有研究缺乏对国际舆情在主题和情感维度的动态探索。另一方面，以新冠疫情为背景的舆情分析研究，缺乏国际主题-情感的动态演化分析。本文以突发公共卫生事件为背景，关注国际媒体对于该事件报道的主题和情感维度随时间的变化。提出了国际舆情演化分析框架，采用主题模型，生成更符合当下舆论风向的主题分类，证明了图神经网络模型在情感分类任务的优越性。结合报道的时间和空间信息进行国际舆情时空演化分析，探索世界各国舆情演化之间的内在关联，为相关部门的政策制定提供理论支撑。

## 3 研究方法

本文以突发公共卫生事件为背景，构建国际舆情演化的研究框架(见图1)。

1）数据采集与预处理：获得国际媒体关于“COVID-19”的报道，进行数据预处理。

2）主题-情感信息挖掘：采用BERT-LDA挖掘舆情数据主题信息，由图神经网络模型挖掘舆情数据的情感倾向，并证明模型的优越性。

3）舆情周期划分及时空演化分析：根据舆情演化三阶段模型，划分为初始传播、迅速扩散、消退。结合报道时空信息和主题-情感信息进行国际舆情时空演化分析及现象挖掘。

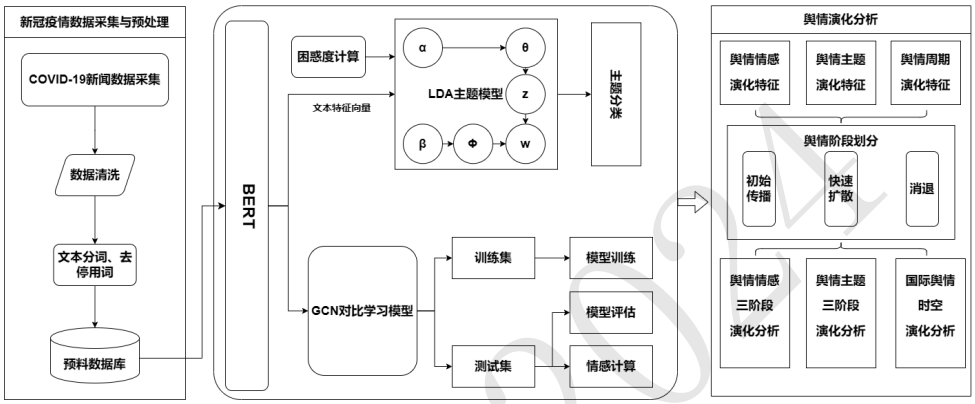



图 1: 国际舆情演化分析研究框架

### 3.1 图神经网络情感分类模型

本文采用图神经网络情感分析模型(见图2)对舆情文本进行情感分析。传统的词向量模型适合对短句和简单的句子进行分析。针对长文本情感分类，还应考虑上下文本及单词与单词之间的关系的问题。首先，利用BERT预训练模型充分挖掘文本上下文特征信息，生成舆情文本的特征向量$V$(见式1)，将所有文本的特征向量作为舆情图的节点集合：

$$\mathbf{V} = BERT(t) = [\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_n] \tag{1}$$

通过余弦距离计算舆情图节点集合中不同节点之间的语义相似度(见式2)，当两节点之间距离在特定范围内时，构建其两点语义关系的边(见式3)。在构建的初始舆情图的基础上引入三层图神经网络模块学习节点嵌入进行特征融合，输出考虑相似文本节点之间相互作用的特征信息(见式4)(Zeng et al., 2023)：

$$\text{CosineDistance} = \frac{\mathrm{V} \cdot \mathrm{V}'}{|\mathrm{V}| \times |\mathrm{V}'|} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{V_i} \times \mathrm{V}'_{\mathrm{i}})}{\sqrt{\sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{V_i})^2} \times \sqrt{\sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{V}'_{\mathrm{i}})^2}} \tag{2}$$

$$E_{ij} = \begin{cases} 1, 0 <= \text{ ConsineDistance }_{ij} <= 0.2 \\ 0, \text{ other} \end{cases} \tag{3}$$

$$v_i^{(l+1)} = \sigma\left(\sum_{j \in N_i} \frac{1}{c_{ij}} v_j^{(l)} w^{(l)} + b^{(l)}\right) \tag{4}$$


(c) 2024 中国中文信息学会计算语言学专业委员会

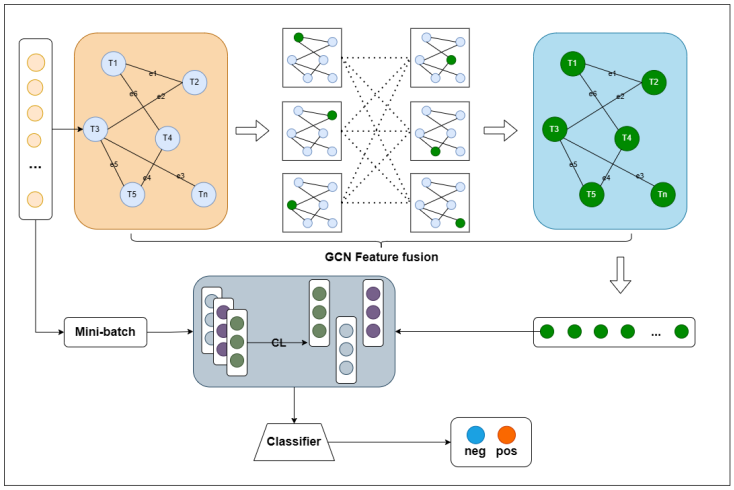



图 2: 基于图神经网络的情感分类模型

其中，$t$为舆情文本数据，$V$为舆情文本的特征向量，$n$为向量维度，$V_i^{'}$为GCN生成的融合节点特征，$E_{ij}$为语义关系边，$v_j^{(l)}$表示在第l层节点j的特征，$\sigma$表示非线性变换，$c_{ij}$为归一化因子，$N_i$表示节点i和节点i的邻居，$W^{(l)}$表示第l层的权值，$b^{(l)}$表示第l层的截距。

为了提高模型在情感分析任务中的泛化能力，本文引入对比损失函数，基于源数据集的情感标签信息，对隐藏向量进行对比学习。给定小批量文本的隐藏向量$H = \{V_i\}_{i-1}^{N_b}$，对于一个$V_i \in H$，如果$V_t$和$V_i$有相同的情感标签，即$y_t = y_i$则$V_i$被认为是$V_t$的正例，而其他例子$V_x \in H$被认为是负例。计算所有正例$(V_t, V_i)$和$(V_i, V_t)$的对比损失(见式5，6)(Liang et al., 2022)。将最终的特征向量输入到分类器中进行分类，使用交叉熵损失函数作为分类器损失函数:

$$\mathcal{L}_{cl} = \frac{1}{N_b}\Sigma_{V_i \epsilon H} l(V_i) \tag{5}$$

$$l(V_i) = \log\left(\frac{\sum_{t=1}^{N_b} 1_{[t\neq i]} 1_{[y_t=y_i]} exp\left(\frac{sim(V_t,V_i)}{\tau_s}\right)}{\sum_{x=1}^{N_b} 1_{[x\neq i]} exp\left(\frac{sim(V_x,V_i)}{\tau_s}\right)}\right)^{-1} \tag{6}$$

其中$1_{[\mathbf{t}\neq\mathbf{i}]}\epsilon(0,1)$为指示函数，$\mathrm{sim}(\mathrm{V_t}, \mathrm{V_i})$计算向量$V_t$和$V_i$的余弦相似度。

图神经网络模型利用BERT预训练模型进行文本的特征编码，增大模型提取特征和融合上下文特征的能力。由于舆论文本的情感分类任务相同类别的不同文本之间存在一定的关联，通过引入图神经网络融合不同文本的特征表示，设置融合模块通过学习相邻节点的信息来更新节点自身的特征表示，生成考虑相似文本节点之间相互作用的新特征，可以有效的增加舆情情感分类的准确性。为增加模型的泛化能力充分利用文本相似性，引入对比学习模块更好的将模型学习到的类别特征延伸至不同的目标数据，进一步提高了分类的准确性。本文利用BERT预训练模型，引入图神经网络特征融合及对比学习模块，有效的利用舆情文本的相似性增强舆论情感特征提取能力，在舆论情感分类任务中具有较好的性能。

### 3.2 舆论主题聚类

为了剖析全球新闻媒体的舆情演化趋势，探究研究期间新闻媒体讨论的主要话题，本文构建BERT-LDA模型进行主题细粒度分析。针对LDA模型无法有效提取上下文信息的问题，本文通过BERT预训练模型提取文本特征，并与LDA模型融合进行主题聚类分析，实现报道主题与报道上下文结合。困惑度计算常用来评价LDA模型的预测和泛化能力，困惑度越低表示模型性能越好(蔡政群et al., 2022)。实验中，困惑度计算(见式7)确定主题最优数$k$。

$$Perplexity(D) = exp\left[\frac{\sum_{d=1}^{M} \log p(w_d)}{\sum_{d=1}^{M} N_d}\right] \tag{7}$$


(c) 2024 中国中文信息学会计算语言学专业委员会

其中$D$为文档所有词的集，$M$为文档数量，$W_d$为文档$D$中的词，$N_d$为每个文档中$D$的词数，$p(w_d)$为文档中词出现的概率。

## 4 实证分析

### 4.1 数据收集与预处理

在深入挖掘新闻文本内部蕴含的主题-情感特征信息之前，本文选取谷歌新闻数据库GDELT (The Global Database of Events, Language, and Tone)设置关键检索字段“V1 THEMES”为“COVID-19”，本文认为在该事件爆发初期存在着较多的因无知所造成的影响，这其中缺乏人们对待这个事件理性的认知以至于存在较大的不确定性，最终选取实验时间区间为2022年7月1日 2022年8月31日，获取新闻来源、新闻获取链接、发布位置信息等属性信息，利用爬虫爬取指定链接的新闻报道。经数据预处理获取新闻事件120,679条，使用大规模语言模型对数据集进行情感倾向标注，其中消极为0积极为1，具体来说，本文在大模型的标注阶段使用的为通义千问作为数据标注的大模型，在标注阶段设置的prompt是以ICIO框架作为基础框架对其中的提示条件例如描述具体任务、任务的背景信息引导模型做出更好的回复、输入数据的格式以及输出数据的格式进行设置。为了验证大模型标签的有效性，我们采用人工验证选取部分数据一组为500条，邀请两位了解新冠肺炎事件的团队成员对两组数据进行标注，将两组人工标记数据分别与大模型标记数据进行对照试验，度量标准为kappa（k）系数，其中结果分别为0.69、0.73，整体一致性较高，证明数据集标注的有效性。

### 4.2 舆论情感演化分析

#### 4.2.1 情感分类模型评估

运用图神经网络模型对舆论进行情感分类，结合自身数据集特性模型参数设置见表1。为验证提出的情感分类模型在该任务上的优越性，选取四个对照实验证明模型的有效性，选取精确率P(Precision)、召回率R(Recall)、F1值(F1-Score)和准确率(Accuracy)作为评估指标。实验对比基本模型选择TF-SVM(王新宇, 2016)、LSTM(王楠and 吕欣隆, 2021)、BERT-SVM(严佩敏and 唐婉琪, 2020)、BERT-AAD(Ryu et al., 2022)，对比实验结果见表2。

| 参数名称 | 参数说明 |
|---|---|
| GCN学习率 | 0.001 |
| GCN隐藏层数 | 3 |
| *Batchsize* | 256 |
| 预训练次数 | 5 |
| 损失函数 | 交叉熵损失函数 |
| 函数优化 | Adam算法 |

表 1: 模型参数设置

| 实验模型 | 模型对比侧重 | *P* | *R* | *F1* | *Acc* |
|---|---|---|---|---|---|
| TF-SVM | 词袋模型不能更好的提取文本特征 | 0.5312 | 0.5518 | 0.5366 | 0.5523 |
| LSTM | 无预训练模型且为单向记忆网络 | 0.6183 | 0.6337 | 0.6119 | 0.6350 |
| BERT-SVM | 未考虑相似文本之间的相互影响 | 0.6512 | 0.6734 | 0.6672 | 0.6833 |
| BERT-AAD | 未考虑相似文本之间的相互影响 | 0.7026 | 0.7219 | 0.7013 | 0.7230 |
| BGCN | 模型泛化性较差 | 0.7405 | 0.7463 | 0.7311 | 0.7387 |
| BGCN-CL | | **0.7857** | **0.7880** | **0.7826** | **0.7973** |

表 2: 模型评估对比

实验结果表明，图神经网络情感分类模型在性能上优于其他模型。对比传统的词袋模型以及深度学习模型，引入BERT预训练模型可以更好的提取文本的特征用于分类任务。BGCN-CL模型在F1指数上比BERT-AAD模型有着较明显的提高，说明考虑相似文本之间的相互影响


(c) 2024 中国中文信息学会计算语言学专业委员会

是必要的。在去除CL对比学习模块后模型性能有所下降，证明了对比学习模块在模型中的有效性，可以有效的提升模型学习不同类别的特征，从而获得更好的模型泛化能力。综上，本文提出的模型可以有效的对舆情进行情感分析，情感倾向预测结果样例见表3

| 文本ID | 文本部分内容 | 标签结果 |
|---|---|---|
| 20220706073000-544 | From digital payments to video banking, the way consumers conduct basic financial transactions has changed substantially throughout the past decade, and the COVID-19 pandemic has only accelerated the trend toward digital banking. | NEG |
| 20220803133000-750 | CBC News LoadedBritish ColumbiaB.C.opens up immunizations to children 4 and under in next phase of COVID-19 fightInfants, toddlers and pre-schoolers can now get appointments for the first of two recommended Moderna Spikevax COVID-19 vaccine doses. Social SharingChildren between six months and four years old can now get appointments for the Moderna vaccine David P. | POS |
| 20220831134500-741 | Get our free mobile app According to their website, the program which launched this past January has not been funded by the Federal Government to continue. Because Congress hasn't provided additional funding to replenish the nation's stockpile,The good news is that you still have time to place your order for the free kits before Friday amp; apos; s final day. HOW TO GET FREE COVID-19 TESTING KITSEach household can order 8 free COVID-19 at-home testing kits for free. | POS |

表 3: 预测结果样例

### 4.2.2 舆论情感态势分析

针对舆情新闻数据及同一时间新增病例数目进行计量分析(见图3)，由图可知舆情报道数量和感染病例数量有着直接关系，即随着当日确诊人数的增多，

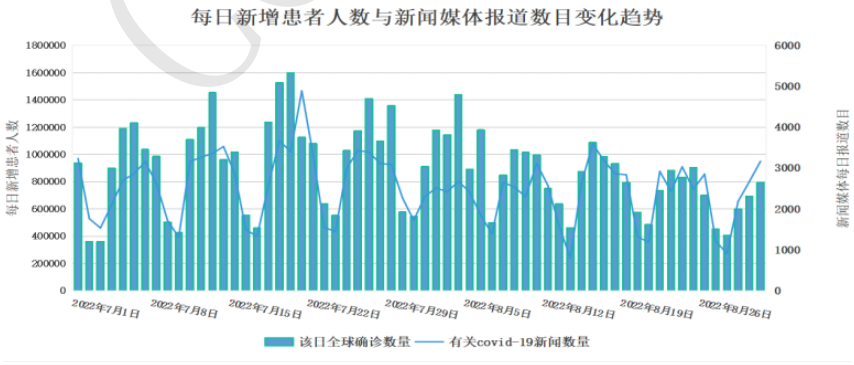



图 3: 新闻报道与新增病例数目统计

有关舆情的新闻报道数量也会有明显的增长趋势。因此，感染数目的变化会导致新闻媒体上关于舆情的报道数量变化(Burkholder and Toole, 1995)。图中曲线变化体现了从开始到爆发再到消退的三阶段现象，这也得益于舆情信息公开及在爆发初期政府进行策略优化，进一步消除公众的不知情恐慌与无端猜测，遏制负面情绪的增长(张冬and 魏俊斌, 2021)。


(c) 2024 中国中文信息学会计算语言学专业委员会

将新闻中的消极词汇和积极词汇进行统计，计算在报道中的情绪占比即Positive Score、Negative Score以日平均为单位进行可视化分析(见图4)。由图可知Positive Score与Negative Score变化趋势呈现明显的负相关关系，当新闻媒体报道中出现的带有积极情绪的词汇占比增加，报道所用的消极词汇的数量就会减少，反之亦然。




图 4: 消极词汇和积极词汇占比时序变化

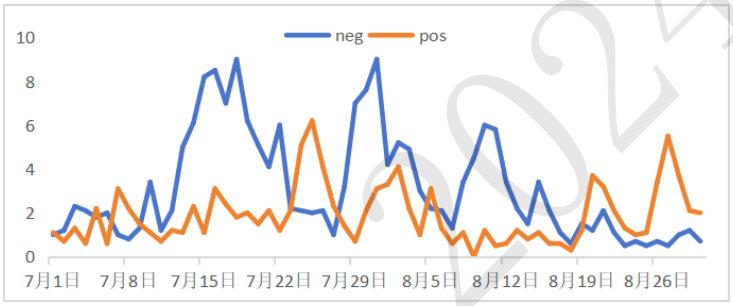



图 5: 情感倾向时序演化图

采用情感分析模型研究国际媒体情感演化过程，获得情感随时间变化趋势(见图5)。由图可知全球新闻媒体的新闻报道消极情感高于积极情感，且在2022年7月18日和2022年7月31日新闻报道的消极情绪达到顶峰。媒体的积极情感整体表现为局部波动，积极情感与消极情感大体呈现负相关的特点。舆情生命周期划分及特点见表4。

| 阶段 | 时间段 | 特点 |
|---|---|---|
| 初始传播 | 7月1日-7月13日 | 消极情绪呈上升趋势，积极情绪平稳而消极情绪有小幅度震荡 |
| 快速扩散 | 7月13日-8月12日 | 消极和积极情绪均处于大幅度震动，消极情绪波动大于积极情绪，并存在负相关 |
| 消退 | 8月12日-8月31日 | 情绪态度趋于稳定，积极情绪逐渐增加 |

表 4: 舆情生命周期划分及特点

初始传播阶段，在突发公共卫生事件的背景下，全球经济形势不容乐观“印度洋岛国斯里兰卡总理维克勒马辛哈宣称斯里兰卡已经破产。他告诉国会，国家已经陷入深度衰退，食品、燃


(c) 2024 中国中文信息学会计算语言学专业委员会

料和药品的严重短缺将继续存在。”此事件的发生得到了全球各界的高度关注，使得群众对于突发公共卫生事件平稳后的经济状况感到担忧，但因信息传播速度影响，公众在该阶段对该事件的态度相对平稳。

迅速扩散阶段，突发公共卫生事件的出现使得全球经济出现新的挑战，中国作为全球第二大经济体深受影响，使得中国在国际上的政治外交变得困难。中美双边关系影响，导致欧美国家对中国的报道相对负面。美国众议院议长佩洛西不顾中方坚决反对，执意在时局动荡的情况下“访台”，严重践踏一个中国原则，损害我国的利益。中国始终坚持主权和领土完整原则，由于中美关系的影响可能导致欧洲和美国对中国的报道相对负面。在该时段内，国际关于突发公共卫生事件的讨论多以中国为主，即中国主流媒体为主要来源。随着国家及主流媒体引导在短时间中出现积极情绪主导的状态。以美为首的西方发达国家因新冠病毒限制中国外交，强行干预国家内政及大肆宣扬突发公共卫生事件为中国病毒导致国际舆论恶化。由于外交和病毒等多方原因造成围观热度和舆论压力，使得公众情绪在恐慌和抱怨中逐渐转向以负面情绪为主导。

消退阶段，由于国家增加对突发公共卫生事件舆情的管控以及猴痘疫情逐渐蔓延，人们将关注的重心逐渐转移到其他相关症状上，虽然有关该事件的舆论热度逐渐消退，公众情绪呈现积极向上的状态，但负面情绪仍有波动性起伏。

### 4.3 舆论主题演化分析

为直观分析舆情热点制作新闻热点词云图(见图6)。词云图展现国际更多关注卫生防控及病毒疫苗的研发，带有明显情感倾向的词语出现较少，“epidemic”、“N95”、“prevention”、“vaccine”、“control”等关键词表明国际讨论重心着眼于如何预防病毒以及治疗病毒、如何抗击病毒等卫生防控角度，反映该事件舆情态势目前较于稳定。本文意在研究舆论主题随时间的转变，探究各个阶段主题的动态变化获取国际舆论风

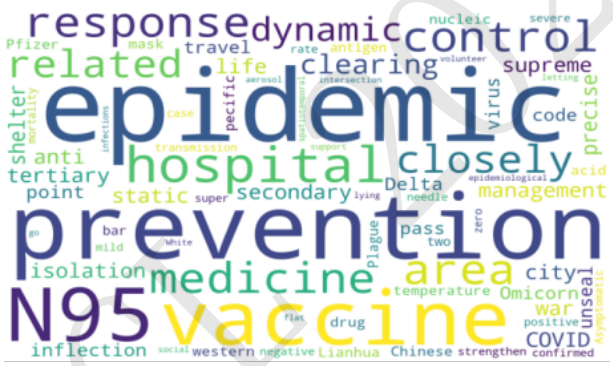



图 6: 热点词云图

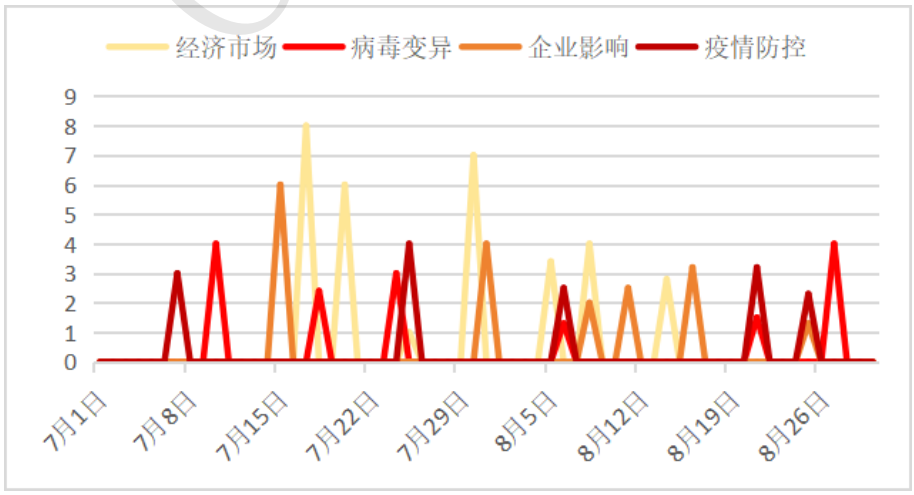



图 7: 主题时序演化图


(c) 2024 中国中文信息学会计算语言学专业委员会

向。基于困惑度计算确定最优主题数并进行主题分类得到频率均值最高的四组主题分别为“经济市场”、“病毒变异”、“企业影响”、“疫情防控”，由潜在主题中出现词汇概率排位较高的十个特征词来表示舆论主题，见表5-7。计算不同阶段下新闻报道主题与上述四组主题的相关性，统计每个主题出现频率随时间的动态演变(见图7)。

初始传播阶段，舆论倾向病毒变异和病毒防控两个部分。病毒并没有大规模爆发，讨论如何应对新一轮的卫生防控成为首要问题。结合情感倾向结果分析出媒体舆论情感相对稳定，正负情感均有波动但幅度较小。政府应加强病毒知识科普，完善公众防护措施控制舆论风向。

| 主题 | 关键词 |
|---|---|
| 病毒变异 | ’covid’,’virus’,’health’,’hospital’,’epidemic’,’N95’,’delta’, ’omicorn’, ’policy’, ’public’ |
| 疫情防控 | ’epidemic’,’prevention’,’N95’,’covid’,’health’,’sports’, ’vaccine’,’quarantine’,’public’,’positive’ |

表 5: “初始传播阶段主题-词”

快速扩散阶段，随着病毒的大规模扩散对社会所造成的经济损失不断增加，股票下跌、企业盈亏的现象增加导致国际对突发公共卫生事件下的经济市场及企业影响的讨论增加。病毒变异及防控的讨论随着扩散规模的严重也显得尤为重要。主题词中的“negative”，“decline”反映出世界经济形势较差，结合情感倾向结果分析媒体舆论倾向较为消极且情感波动较大。政府应积极制定经济相关政策，引导主流媒体对该事件积极信息的报道传达主流价值观，减少公众因消极信息而带来的恐慌提升政府公信力。

| 主题 | 关键词 |
|---|---|
| 经济市场 | ’pneumonia’,’amp’,’covid’,’business’,’market’,’china’, ’area’,’policy’,’republic’,’negative’ |
| 病毒变异 | ’covid’,’virus’,’health’,’hospital’,’epidemic’,’N95’,’delta’, ’omicorn’, ’policy’, ’public’ |
| 企业影响 | ’enterprise’,’Biden’,’business’,’market’,’policy’,’response’, ’decline’,’amp’,’pandemic’,’evolution’ |
| 疫情防控 | ’epidemic’,’prevention’,’N95’,’covid’,’health’,’sports’, ’vaccine’,’quarantine’,’public’,’positive’ |

表 6: “快速扩散阶段主题-词”

消退阶段，媒体更加关注大规模卫生健康问题对企业经济的影响以及防控措施和病毒知识的科普。结合情感倾向结果分析媒体舆论情感整体较为积极，源于政府积极普及防疫措施和疫苗注射并取得了较好的效果，有效的减小该事件所造成的危害。通过主流媒体宣传防疫成果引导社会舆论倾向，使得公众对待该事件的态度出现由消极到积极的转变。

| 主题 | 关键词 |
|---|---|
| 病毒变异 | ’covid’,’virus’,’health’,’hospital’,’epidemic’,’N95’,’delta’, ’omicorn’, ’policy’, ’public’ |
| 企业影响 | ’enterprise’,’Biden’,’business’,’market’,’policy’,’response’, ’decline’,’amp’,’pandemic’,’evolution’ |
| 疫情防控 | ’epidemic’,’prevention’,’N95’,’covid’,’health’,’sports’, ’vaccine’,’quarantine’,’public’,’positive’ |

表 7: “消退阶段主题-词”


(c) 2024 中国中文信息学会计算语言学专业委员会

### 4.4 舆情主题-情感时空演化分析

为动态观察各国舆论主题随时间变化规律，根据上文划分的阶段，对不同阶段各个国家的主题向量进行聚类按照时间序列进行可视化展示，绘制全球地图其中国家颜色与图7主题颜色对应(见图8)。




图 8: 主题-情感时空演化图

由图8可知，突发公共卫生事件舆情在国际上具有以舆情处于初始爆发阶段国家为中心，向周围国家放射性传播舆情态势的特点。以中国即周边国家为例，在中国周边国家处于舆情的迅速扩散阶段时，中国境内正在处于初始传播阶段。随着舆情在周边国家的发展，中国由初始传播阶段进入到迅速扩散阶段，政府的积极治理使得中国舆情由迅速扩散到消退的转变。对比多个国家发现中国在由初期传播到消退的转变明显，反映出中国虽作为发展中国家，但对于突发卫生事件的防控及认真态度站在世界前列，中国各级政府积极响应中央号召，果断实施相应政策，极大缓解因疫情造成的经济损失及公众恐慌，为世界其他国家的卫生防控提供了一套重要的管理模式，为世界公共卫生安全做出了重要的贡献。

## 5 总结与展望

本文针对与“COVID-19”相关新闻报道构建了从“国际舆情数据收集及预处理”、“舆情情感分类”、“舆情主题聚类”、“国际舆情周期演化分析”的全过程研究框架，揭示了突发性公共卫生事件在全球各时间段的发展态势，通过对比实验证实了模型的优越性。以“初始传播、迅速扩散、消退”三个阶段在时间维度上进行研究，发现在初始传播和消退阶段的新闻报道在情感和讨论主题上处于相对稳定的状态，而在快速扩散阶段由于确诊数量的不断增加，也增大了各国媒体对该事件的担忧程度，从而使得各国新闻的情感和讨论主题更加偏向于消极的态度。在全球空间维度研究中，针对各国媒体的舆论主题和情感倾向，结合三阶段时间维度动态观察每个阶段国家之间的相互影响。分析实验结果得出结论，突发性公共卫生事件背景下各国之间存在舆情态势放射性传播的现象，当目标国家出现并爆发大规模疫情后随着时间的推移，邻国也会慢慢的由相对平稳的初始传播阶段转入快速扩散阶段。该发现有利于各国政府在邻国出现疫情问题后，及时地制定相关外来人员入境政策以防止国内大规模疫情的爆发。

本文存在一些不足，如选取的数据时间跨度相对较短，无法更加多样、更加深层的挖掘舆情演化态势。数据类型局限于报道文本，较为单一。今后的研究会着重考虑增加数据类型，引入多个模态的数据更加深层的挖掘舆情演化。扩大研究数据集深层挖掘舆情时间维度信息，为相关部门提供更全面的舆情现象分析及理论支撑。

## 参考文献

Lu An, Chuanming Yu, Xia Lin, Tingyao Du, Liqin Zhou, and Gang Li. 2018. Topical evolution patterns and temporal trends of microblogs on public health emergencies: an exploratory study of ebola on twitter and weibo. *Online information review*, 42(6):821–846.

Brent T Burkholder and Michael J Toole. 1995. Evolution of complex disasters. *The Lancet*, 346(8981):1012–1015.

Lucien G Canton. 2006. San francisco 1906 and 2006: An emergency management perspective. *Earthquake spectra*, 22(2_suppl):159–182.


(c) 2024 中国中文信息学会计算语言学专业委员会

Quan Cheng, Jianhua Kang, and Minwang Lin. 2021. Understanding the evolution of government attention in response to covid-19 in china: A topic modeling approach. In *Healthcare*, volume 9, page 898. MDPI.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Bo Huang, Yan Yang, Amjad Mahmood, and Hongjun Wang. 2012. Microblog topic detection based on lda model and single-pass clustering. In *International Conference on Rough Sets and Current Trends in Computing*, pages 166–171. Springer.

Bin Liang, Qinlin Zhu, Xiang Li, Min Yang, Lin Gui, Yulan He, and Ruifeng Xu. 2022. Jointcl: A joint contrastive learning framework for zero-shot stance detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 81–91. Association for Computational Linguistics.

Cristian R Machuca, Cristian Gallardo, and Renato M Toasa. 2021. Twitter sentiment analysis on coronavirus: Machine learning approach. In *Journal of Physics: Conference Series*, volume 1828, page 012104. IOP Publishing.

Minho Ryu, Geonseok Lee, and Kichun Lee. 2022. Knowledge distillation for bert unsupervised domain adaptation. *Knowledge and Information Systems*, 64(11):3113–3128.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Ziming Zeng, Shouqiang Sun, and Qingqing Li. 2023. Multimodal negative sentiment recognition of online public opinion on public health emergencies based on graph convolutional networks and ensemble learning. *Information Processing & Management*, 60(4):103378.

严佩敏and 唐婉琪. 2020. 基于改进bert的中文文本分类. 工业控制计算机, 33(7):4.

史伟, 薛广聪, and 何绍义. 2022. 情感视角下的网络舆情研究综述. 图书情报知识, 39(1):105–118.

史波. 2010. 公共危机事件网络舆情应对机制及策略研究. 情报理论与实践, (7):4.

张冬and 魏俊斌. 2021. 情感驱动下主流媒体疫情信息数据分析与话语引导策略. 图书情报工作, 14.

曾子明and 万品玉. 2018. 融合演化特征的公共安全事件微博情感分析. 情报科学, 36(12):3–8.

曾子明and 陈思语. 2023. 基于lda与bert-bilstm-attention模型的突发公共卫生事件网络舆情演化分析. 情报理论与实践, 46(9):158–166.

潘崇霞. 2011. 网络舆情演化的阶段分析. 计算机与现代化, (10):203–206.

王新宇. 2016. 基于情感词典与机器学习的旅游网络评价情感分析研究. 计算机与数字工程, 44(4):578–582.

王楠and 吕欣隆. 2021. Covid-19 疫情下政务媒体情感传播特征及用户情感体验研究. 情报探索, 1(2):1.

甘宇祥, 王亚博, 薛均晓, 张若琪, 许书宁, and 郭毅博. 2021. 基于情感特征的新冠肺炎疫情舆情演化分析. 图学学报.

蔡政群, 李诗轩, 吴玉敏, and 杨斯涵. 2022. 基于lda模型的学术次生衍生舆情传播的演化分析. 武汉理工大学学报:信息与管理工程版, 44(3):395–401.

边晓慧and 徐童. 2022. 重大突发公共卫生事件下的公众情感演进分析:基于新冠肺炎疫情的考察. 数据分析与知识发现, (7):128–140.

邓君, 孙绍丹, 王阮, 宋先智, and 李贺. 2020. 基于word2vec 和svm 的微博舆情情感演化分析. 情报理论与实践, 43(8):112–119.


(c) 2024 中国中文信息学会计算语言学专业委员会

# 面向社交媒体多特征增强的药物不良反应检测

**李浩，邱云志，林鸿飞***
大连理工大学，计算机科学与技术学院，大连，116024
(lihao，yzqiu)@mail.dlut.edu.cn,(hflin)@dlut.edu.cn

## 摘要

社交媒体是药物不良反应(ADR) 检测的重要途径之一。本文提出一个基于社交媒体的药物不良反应检测模型DMFE，以全面捕捉患者对药物使用的反馈信息。与传统的文本检测相比，社交媒体数据中通常会有语法不规范与单词拼写错误的问题。本文提取出社交媒体数据的抽象语义表示（AMR）使用图注意力网络（GAT）学习抽象语义特征提高模型对语义信息的理解，使用字符级卷积神经网络（charCNN）捕获字符特征以减少单词拼写错误带来的影响。此外，本文使用提示学习的方法融入MedDRA药物不良反应领域关键词进一步增强模型对领域知识的理解能力。经实验评估，本文模型DMFE在CADEC、TwiMed两个数据集上F1值与基线模型相比取得最优效果。

**关键词：** 社交媒体 ； 药物不良反应 ； 文本检测

# Adverse drug reaction detection with multi-feature enhancement for social media

**Hao Li，Yunzhi Qiu，Hongfei Lin***
School of Computer Science and Technology,Dalian University of Technology,Dalian,116024
(lihao，yzqiu)@mail.dlut.edu.cn,(hflin)@dlut.edu.cn

## Abstract

Social media is one of the important ways to detect adverse drug reactions (ADR). This paper proposes a social media-based adverse drug reaction detection model DMFE to capture patients' feedback on drug use. Compared with traditional text detection, social media data often contains grammatical irregularities and word spelling errors. In this paper, abstract semantic representation (AMR) of social media data is extracted, abstract semantic features are learned by graph attention network (GAT) to improve the model's understanding of semantic information, and character level convolutional neural network (charCNN) is used to capture character features to reduce the impact of word spelling errors. In addition, this paper uses the method of prompt learning to integrate key words in the field of MedDRA adverse drug reactions to further enhance the model's understanding of domain knowledge. According to the experimental evaluation, the F1 value of the DMFE model in this paper achieves the optimal effect compared with the baseline model on the CADEC and TwiMed datasets.

**Keywords:** Social media , Adverse drug reaction , Text detection

* 通讯作者

第二十三届中国计算语言学大会论文集，第515页-第525页，太原，中国，2024年7月25日至28日。
卷1：主会论文
(c) 2024 中国中文信息学会计算语言学专业委员会

## 1 引言

药物不良反应（Adverse Drug Reaction，简称ADR）是指正常剂量的药物用于预防、诊断、治疗疾病或调节生理机能时出现有害的、与用药目的无关的反应(National Center for ADR Monitoring，China, 2023)。这些反应是预期之外的，会严重影响患者的生理状态与健康状况，甚至会有致命的后果。因此及时准确地发现潜在的不良反应对于确保药物的安全性至关重要(Marano et al., 2005)。

随着社交媒体的快速发展，大量用户选择在社交媒体平台上分享自己的用药经验，其中包括处方药和非处方药的使用、副作用以及治疗效果，社交媒体中的大量用药信息为在社交媒体中检测药物不良反应提供可能。在社交媒体中检测药物不良反应相较于普通文本检测任务主要有以下的特点：

（1）在社交媒体中，用户经常以非正式的方式分享交流，导致数据语法层面规范性普遍缺失。例如，在句子中省略主语、谓语、连词，或者混用不同时态的动词，这样的语言风格会对检测模型造成理解上的困扰。

（2）使用社交媒体过程中，用户追求表达速度导致单词拼写错误、使用缩写的现象十分常见，例如将“their”误写成“there”，用“meddizzy”来代替“medication”和“dizzy”等。另外对词性的不正确使用也是普遍问题，用户可能在不经意间将名词用作动词，将形容词用作名词，例如将“affect”误用为名词形式“effect”，以上都会给模型检测带来不确定性。

（3）社交媒体数据中关于药物不良反应的讨论往往围绕着病痛、药物名称以及不良反应症状等关键词展开。用户在描述某种药物的副作用时可能会提到诸如“headache”、“nausea”、“insomnia”等病痛相关的单词；如“aspirin”、“antibiotic”等药物名称相关单词；如“skin allergy”、“rapid heartbeat”等不良反应的症状相关单词，增强模型对领域知识的理解有助于提高模型在社交媒体中药物不良反应检测的能力。

针对上述问题，本文提出一个融合药物不良反应领域知识，字符级别信息以及抽象语义表示（AMR）的模型：面向社交媒体多特征增强的药物不良反应检测（Adverse drug reaction detection with multi-feature enhancement for social media，简称DMFE），本文工作如下：

（1）对MedDRA数据库进行预处理并完成词级编码以获取药物不良反应领域词典。将社交媒体数据在词典中检索，并将检索到的关键词使用提示学习的方式与社交媒体数据共同融入Prompt模板。

（2）使用字符级别的卷积神经网络（charCNN）捕捉社交媒体数据的字符级别特征。卷积神经网络可以捕获到单词内外部特征学习到不同单词之间的相似性从而解决单词字母拼写错误问题和词性使用错误的问题，意思相近但词性不同的单词之间通常只在前后缀有区分。

（3）社交媒体数据的口语化往往导致语法缺失，传统使用句法依赖树增强模型理解语义的方式不再适用。本文使用图注意力网络（GAT）学习抽象语义表示（AMR）的有效信息，提高模型对语义信息的理解。

## 2 相关工作

### 2.1 文本的分类检测

随着深度学习的发展，基于深度学习的模型应用到文本检测任务上取得不错的效果。Chen (2015)提出的基于卷积神经网络的TextCNN模型，首次将CNN应用到文本分类上。Zhang et al. (2015)提出charCNN，将句子离散化为字符级别表示并使用CNN提取字符级信息，字符级信息可以使模型自然而然学会拼写错误和表情符号等异常的字符组合。随着研究的进一步推进，预训练语言模型被应用到自然语言处理的任务中，Sun et al. (2019)研究了Bert在文本分类任务上的不同微调方法，包括文本的预处理、逐层学习率和灾难性遗忘等问题，在文本分类任务上取得了较大的性能改进。

### 2.2 药物不良反应检测

药物不良反应检测任务早期研究工作集中在使用规则匹配以及机器学习的方法从生物医学文献和临床报告中进行检测(Benton et al., 2011; Kuhn et al., 2010; Yates and Goharian, 2013;

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

Patki et al., 2014)。尽管它们在一定程度上取得了不错的效果，这些方法严重依赖于领域知识库与手工特征制作，需要大量专业生物医学领域人员的介入，给检测工作带来额外成本。近年来，深度学习也广泛应用于药物不良反应检测，Huynh et al. (2016)提出了两种新的卷积神经网络模型CRNN、CNNA用于药物不良反应的检测。Zhang et al. (2021)提出了ANNSA模型结合情感注意力机制和对抗学习的检测模型。Alimova and Solovyev (2018)提出交互注意网络(IAN)从用户评论中检测药物不良反应。Wu et al. (2018)提出了一种多头自注意力和分层推文表示的检测方法。Qiu et al. (2023)提出了融入领域关键词的KESDT模型，融入外部领域知识对药物不良反应检测增强，但没有考虑到数据中有大量不规范的语法与单词会造成很大噪声。目前没有检测模型既考虑到社交媒体数据不规范的问题又采用领域关键词的方式提高模型对领域知识的理解能力，本文提出DMFE使用多特征融合的方式解决以上问题，提高药物不良反应的检测能力。

## 3 模型结构

针对在社交媒体中检测药物不良反应的任务特点，本文提出的DMFE模型结构如图1所示，具体如下：（1）输入数据：模型的输入共分为两部分，一部分是需要检测的社交媒体数据，另一部分是从MedDRA预处理编码后的词典中检索到的药物不良反应领域关键词。（2）多特征提取层：分别包括基于Prompt的领域信息与社交媒体数据融合提取，使用charCNN进行的字符级信息提取与使用GAT网络进行的AMR语义信息提取。（3）特征融合层：将特征提取层提取到的三种不同维度特征进行融合。（4）目标检测层：将融合后的特征经过全连接层，然后通过SoftMax函数输出检测结果。

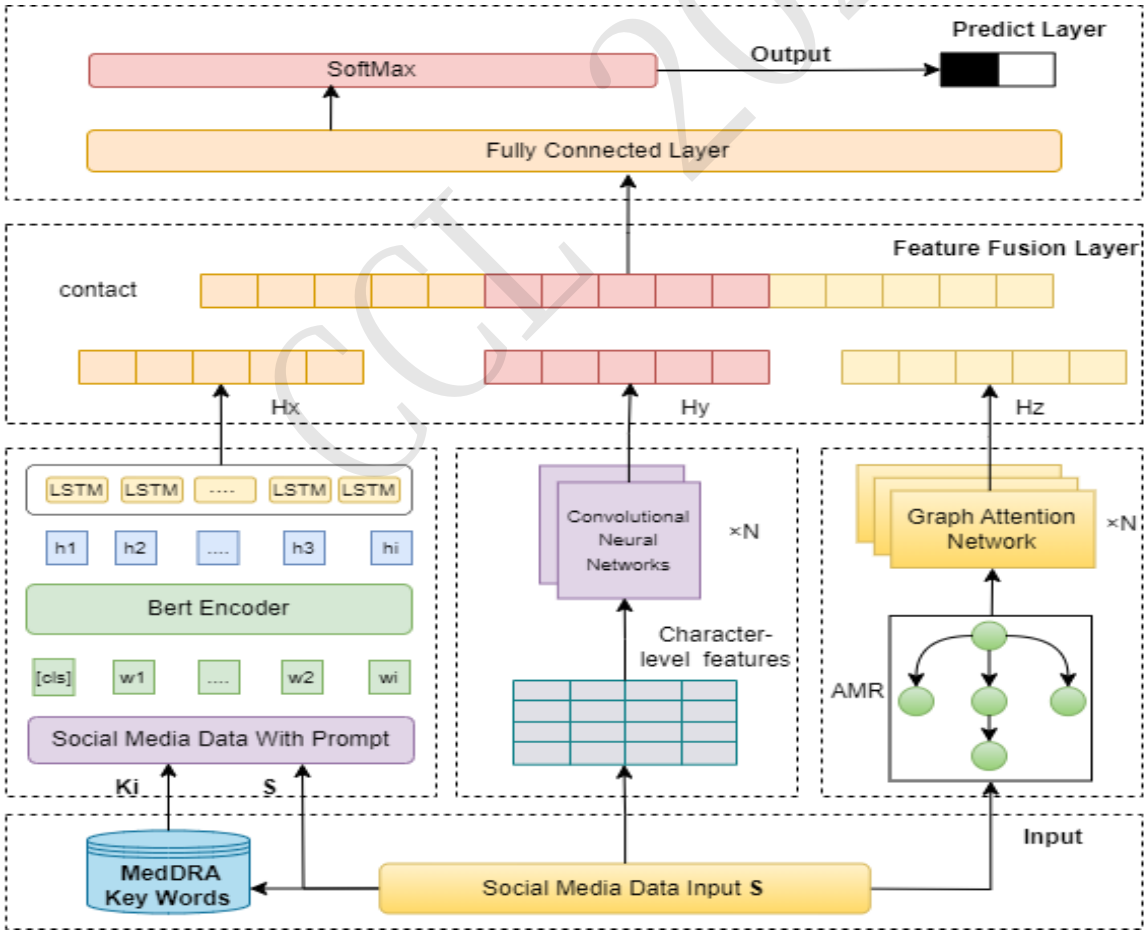



Figure 1: DMFE模型结构图


(c) 2024 中国中文信息学会计算语言学专业委员会

### 3.1 提示学习融合领域知识

#### 3.1.1 MedDRA

MedDRA全称为监管活动医学词典（Medical Dictionary for Regulatory Activities）(Mozzicato, 2009)，是由世界卫生组织主导发展的一种用于药物安全监测的临床术语分类系统。MedDRA广泛应用于药物开发和监管领域，是药物不良事件报告的国际标准，它包括数千个术语用于描述药物治疗过程中患者可能遇到的各种不良事件和副作用。MedDRA收录的医学术语范围包括：体征、症状、疾病、诊断、适应证、各类检查的名称和定性结果等，本文将MedDRA预处理并过滤编码为含有12033个药物不良反应领域关键词的词典。

#### 3.1.2 提示学习

提示学习（Prompt learning）(Liu et al., 2023)可以通过关键信息的提供，激发模型学习理解能力，本文将社交媒体数据在药物不良反应词典中检索到领域关键词$\mathbf{K_i}$，和社交媒体数据$\mathbf{S}$ 按照模板构建提示学习语句，从而融入药物不良反应关键词信息增强模型对领域知识理解。本文最终构建Prompt格式为：The key words:$\mathbf{K_i}$[SEP]Is there adverse drug reaction in the following sentence?[SEP]$\mathbf{S}$ 。

#### 3.1.3 编码层

本文使用Bert(Devlin et al., 2018)作为社交媒体数据与领域关键词构造Prompt之后的编码层。Bert是一种预训练语言模型，其基于Transformer的Encoder架构，通过在大规模文本语料库上进行预训练，学习了丰富的语言表示，其引入了预训练微调的范式，允许在特定任务上微调模型，从而更好地适应特定领域。

#### 3.1.4 特征提取器

通过Prompt融合社交媒体数据与领域关键词之后输入信息长度大大增加，为获取长序列信息的深层次维度特征，本文使用LSTM模型作为编码层词向量的特征提取器。LSTM自循环的权重是不固定的，根据上下文来确定权重值，可以更好地处理长距离依赖问题，LSTM引入了输入门、遗忘门和输出门，输入门控制当前计算的新状态以多大程度更新到记忆单元中；遗忘门控制前一步记忆单元中的信息有多大程度被遗忘掉；输出门控制当前的输出有多大程度上取决于当前的记忆单元，进而实现了对较长序列的语义长短期记忆，具体公式如(1)-(5)所示。

$$i_t = \sigma\left(\boldsymbol{W}_i x_t + \boldsymbol{U}_i h_{t-1} + \boldsymbol{V}_i c_{t-1}\right) \tag{1}$$

$$f_t = \sigma\left(\boldsymbol{W}_f x_t + \boldsymbol{U}_f h_{t-1} + \boldsymbol{V}_f c_{t-1}\right) \tag{2}$$

$$o_t = \sigma(\boldsymbol{W}_o x_t + \boldsymbol{U}_o h_{t-1} + \boldsymbol{V}_o c_{t-1}) \tag{3}$$

$$c_t = f_t \times c_{t-1} + i_t \times \tanh\left(\boldsymbol{W}_c x_t + \boldsymbol{U}_c h_{t-1}\right) \tag{4}$$

$$h_t = o_t \times \tanh\left(c_t\right) \tag{5}$$

$i_t$代表输入门，$f_t$代表遗忘门，$o_t$代表输出门，$h_t$代表隐藏状态节点，$c_t$代表记忆单元，$x_t$是当前时刻t的输入，$\sigma$是logistic激活函数，$\boldsymbol{W}, \boldsymbol{U}, \boldsymbol{V}$是权重矩阵，其中$\boldsymbol{V}_i, \boldsymbol{V}_f, \boldsymbol{V}_o$是对角矩阵。

### 3.2 抽象语义信息提取

#### 3.2.1 AMR抽象语义表示

抽象语义表示(abstract meaning representation，AMR)是句子语义的一种表示方法，可以将社交媒体数据的语义抽象为一个单根有向无环图，其中实词抽象为概念节点，实词之间的关系抽象为带有语义关系标签的有向弧，同时忽略虚词和形态变化体现的较虚的语义（如the、单复数等），本文使用amrlib[0]对社交媒体数据进行解析得到相对应的抽象语义表示。以图2为例，A为AMR原始格式，B为直观抽象语义表示图，使用penman库[1]将抽象语义图格式化，为

[0]https://github.com/bjascob/amrlib/
[1]https://pypi.org/project/Penman/1.3.0/


(c) 2024 中国中文信息学会计算语言学专业委员会

保证AMR语义的词汇信息多样性表示并且加快模型训练推理速度，使用word2vec[2]对AMR语义图中每个节点进行词嵌入表示，无匹配词嵌入使用全0向量表示。

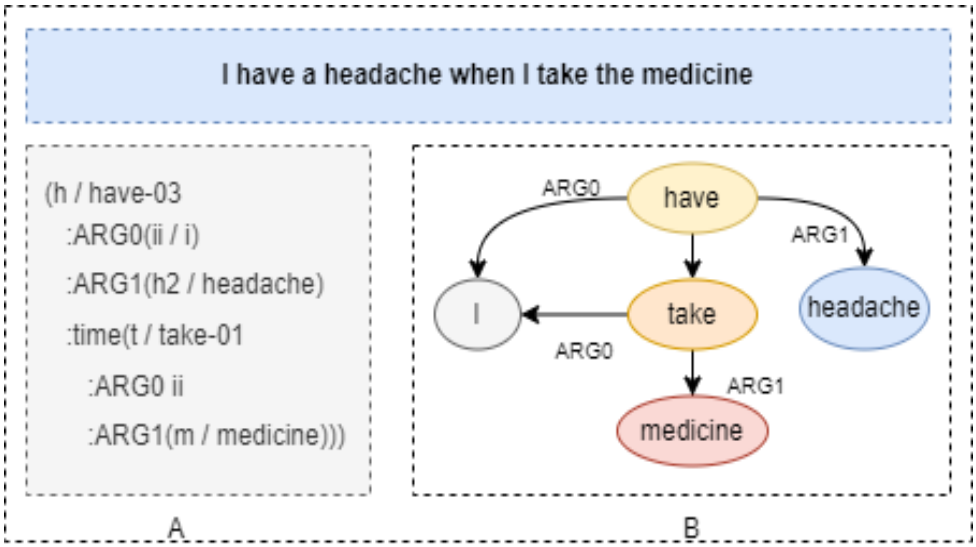



Figure 2: 社交媒体数据的抽象语义表示

### 3.2.2 语义特征学习

图注意力网络（GAT）可以动态捕获社交媒体数据AMR中各节点间的相关性，根据AMR语义图的不同，在聚合邻域特征时为各个节点分配不同权重，因此GAT 比图神经网络（GCN）等基于静态图的方法具有更大的表示空间和更强的表示能力。单个节点更新过程如图3所示，中心节点$h_i$ 扫描所有邻域节点$h_j$对它们的特征向量进行加权求和得到聚合后的特征向量，其中加权求和的权重$\alpha_{ij}$是由$h_i$对每个邻域节点$h_j$ 的注意力分数$e_{ij}$ 经过SoftMax 后得到的，$\mathrm{h}_i \in R^{d_h}$ 是第i 个节点的输入特征向量，$d_h$ 表示隐藏层的特征维度，$\boldsymbol{W} \in \mathrm{R}^{d_h \times d_h}$和$\mathbf{a} \in \mathrm{R}^{2\mathrm{d}_h}$ 表示可学习的参数，$\|$ 表示拼接操作，如公式(6)(7) 所示。

$$e_{ij} = \text{leakyReLU}\left(\mathbf{a}^T \left[\mathbf{W}h_i \parallel \mathbf{W}h_j\right]\right) \tag{6}$$

$$\alpha_{ij} = \text{SoftMax}\left(e_{ij}\right) = \frac{\exp\left(e_{ij}\right)}{\sum_{k \in N_i \cup \{i\}} \exp\left(e_{ik}\right)} \tag{7}$$

聚合时的注意力系数是由中心节点的特征和邻域节点的特征决定的，其随特征的变化自适应地发生改变，因此具有更大的表示空间。如公式(8)所示，$N_i$ 表示节点$h_i$ 的邻域节点集合，$\cup$ 表示合并操作，Relu表示激活函数，聚合邻域节点特征得到中心节点$h_i$ 新的特征表示$\mathrm{h}_\mathrm{i}'$。

$$\mathrm{h}_\mathrm{i}' = \mathrm{Relu}\left(\sum_{\mathrm{j} \in \mathrm{N_i} \cup \{\mathrm{i}\}} \alpha_{\mathrm{ij}} \mathrm{h}_\mathrm{j}\right) \tag{8}$$

为了进一步提高GAT对AMR语义的建模能力，本文设置多个并行的注意力头如公式（9）所示，K 表示注意力头的数量，每个注意力头都有独属于自己的参数，$\mathrm{h}_\mathrm{i}^{\prime(\mathrm{k})}$ 表示第k个注意力头输出的特征向量，$\boldsymbol{W_o} \in \mathrm{R}^{\mathrm{d_h} \times \mathrm{Kd_h}}$为参数矩阵，最后对更新后的所有AMR语义节点做平均池化得到抽象语义表示特征。

$$\mathrm{h}_\mathrm{i}' = \boldsymbol{W_o}\left[\|_{\mathrm{k}=1}^{\mathrm{K}} \mathrm{h}_\mathrm{i}^{\prime(\mathrm{k})}\right] \tag{9}$$

[2]https://github.com/mmihaltz/word2vec-GoogleNews-vectors


(c) 2024 中国中文信息学会计算语言学专业委员会

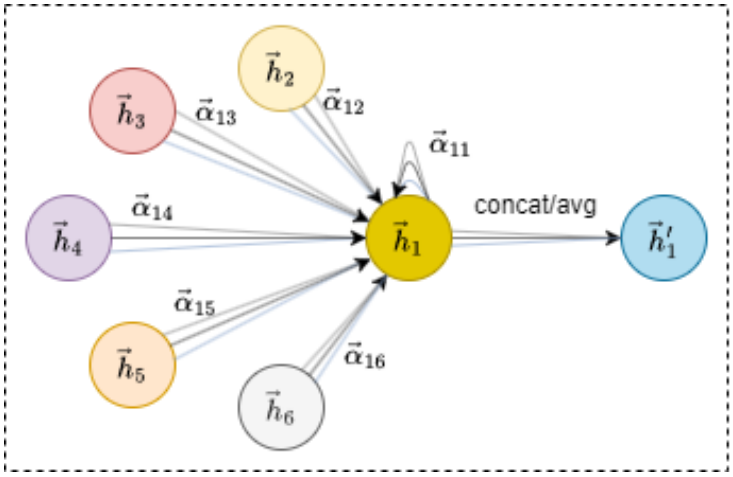



Figure 3: 社交媒体数据AMR语义信息的GAT节点状态更新

### 3.3 字符级信息提取

字符级别特征对解决社交媒体数据中单词表达不规范的问题有很大的优势。通过学习数据整体的字符级表示可纠正单词错误以及学习单词的共性特征，例如模型学习到字符特征后会理解前缀un和dis通常在英文中表示否定，而“their”因为和“there”单词形态相似，二者非常可能由一个误拼写为另一个。Zhang et al. (2021)提出对每个单词使用CNN进行编码提取词嵌入，这样很难利用不同单词之间形态相似关系，只能学习每个单词内部特征，本文使用charCNN(Zhang et al., 2015)直接对字符化的社交媒体数据提取全局字符级别的信息。

本文构建由符号和字母组成的字符表对应使用one-hot表示，共统计了69个字符包含26个英文字母、10个数字、33个其他的字符，当出现不存在字符时使用全0向量表示。将每个社交媒体数据中的字符处理成统一长度，通过多头charCNN和最大池化进行字符级特征提取，然后进行全连接操作生成字符级别的特征向量，过程如图4所示。

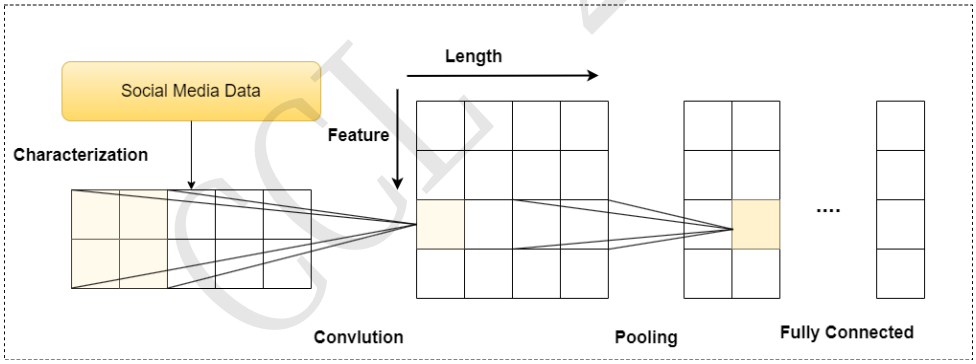



Figure 4: 社交媒体数据字符级别特征提取

### 3.4 多维度特征融合层

特征融合层将基于Prompt的领域信息与社交媒体数据融合特征$H_x$、抽象语义表示特征$H_y$、字符级信息特征$H_z$融合后得到特征融合表示$H_{concat}$，如公式（10）所示。

$$H_{concat} = \text{Concat}\left(H_x, H_y, H_z\right) \tag{10}$$

### 3.5 检测结果输出

#### 3.5.1 损失函数

由于药物不良反应检测任务中存在严重的数据不均衡问题，本模型损失函数采用Focal Loss，相比交叉熵损失函数其对于分类不准确的样本损失没有改变，对于分类准确的样本损失

会变小。整体而言相当于增加了分类不准确样本在损失函数中的权重，具体分类函数如公式（11）所示。

$$L_{fl} = -(1-p_t)^{\gamma}\log(p_t) \tag{11}$$

$\gamma > 0$ 为可调节因子，$p_t$ 反映了预测值与真实标签值的接近程度，$p_t$ 越大说明预测值越接近真实标签值分类越准确。

#### 3.5.2 目标检测层

融合层向量$H_{concat}$经过全连接层通过权重矩阵$\boldsymbol{W}_{\text{out}}$得到最终特征向量$Y$如公式（12）所示，然后通过SoftMax函数进行分类选择较大概率值作为分类结果，如公式（13）所示。

$$Y = \boldsymbol{W}_{\text{out}}\ H_{concat} \tag{12}$$

$$\text{SoftMax}\,(y_i) = \frac{e^{y_i}}{\sum_{i=1}^{n} e^{y_i}} \in (0,1) \tag{13}$$

## 4 实验分析

### 4.1 数据集介绍

为公平评估本文模型DMFE的性能，在两个数据集上对模型进行性能评估。数据集情况如表1所示，下面对使用的两个数据集进行介绍:

（1）CADEC(Karimi et al., 2015)：数据来自于医学论坛AskaPatient用户评论的帖子，每个帖子标有药物、副作用、症状和疾病，将包含副作用的帖子标记为Yes，其余标记为No。

（2）TwiMed(Alvaro et al., 2017)：数据由两部分组成，来自社交媒体平台Twitter的TwiMed-Twitter和来自生物医学文献的TwiMed-PubMed。每个文档都附有疾病、症状、药物及其关系的注释，三种类型的关系为Outcome-negative、Outcome-positive和Reason-to-use。当关系类型为Outcome-positive时，标记为Yes，含有药物不良反应信息；其余标记为No，不含药物不良反应信息。

| 数据集 | ADR | No-ADR | Total | Max_length |
|---|---|---|---|---|
| CADEC | 2478 | 4996 | 7474 | 46 |
| TwiMed | 426 | 1182 | 1608 | 241 |

Table 1: 社交媒体药物不良反应数据集介绍

### 4.2 实验参数

实验使用Pytorch框架，在Ubuntu系统下进行，使用RTX3080进行训练，优化器使用Adam。在两个数据集上进行30个epoch的训练，batchsize设置为32，学习率设置为1×10-5，Dropout设置为默认参数0.1，Bert的隐藏层大小设置为768维，最大句子长度设置为128，损失函数超参数$\gamma$设置为3。

### 4.3 评价指标

本文使用常用的三个评价指标准确率（P）、召回率（R）、F1值来对模型进行评价。其中TP表示药物不良反应正确分类的数据，FN表示含有药物不良反应的数据预测为不含不良反应，FP表示不含不良反应的数据预测为含有药物不良反应，如公式（14）-（16）所示。

$$P = \frac{TP}{TP+FP} \tag{14}$$

$$R = \frac{TP}{TP+FN} \tag{15}$$

$$F1 = \frac{2PR}{P+R} \tag{16}$$


(c) 2024 中国中文信息学会计算语言学专业委员会

### 4.4 DMFE与基线模型对比实验

本文模型DMFE与7个基线模型进行对比实验分析，CADEC与TwiMed数据集上的对比实验结果如表2所示，基线模型介绍如下：

(1)CRNN(Huynh et al., 2016)：一种循环卷积神经网络模型，用于文本分类任务，该模型将卷积层叠加在循环神经网络之上。

(2) HTR+MSA(Wu et al., 2018)：该模型具有分层的多头自注意力机制结构，利用层级表示学习文本的多维度特征，并用于药物不良反应检测。

(3) CNN+corpus(Li et al., 2020)：该方法直接在每个数据集的训练集部分增加了ADE数据，实现对训练语料的扩充，然后在CNN方法上使用扩充后的训练集进行训练。

(4) CNN+transfer(Li et al., 2020)：该模型使用迁移学习方法，将源数据集与目标数据集共同训练，采用CNN学习两部分数据的共享特征，最终使用各自的文本分类器进行预测，得到两部分数据各自的预测结果。

(5) ATL(Li et al., 2020)：结合对抗训练和迁移学习，学习两部分的特征，实现不同数据之间的融合。

(6) ANNSA(Zhang et al., 2021)：融合情感感知的注意力机制的对抗网络模型。

(7) KESDT(Qiu et al., 2023)：融合了药物不良反应领域知识的多层Transformer模型，在Transformer层级之间融入领域关键词。

| 模型 | CADEC | | | TwiMed | | |
|---|---|---|---|---|---|---|
| | **P**(%) | **R**% | **F1**(%) | **P**(%) | **R**(%) | **F1**(%) |
| CRNN | 61.26 | 65.96 | 63.52 | 68.52 | 66.43 | 67.46 |
| HTR + MSA | 60.67 | 61.70 | 61.18 | 66.58 | 63.62 | 65.07 |
| CNN + corpus | 52.75 | 61.28 | 56.69 | 60.51 | 61.50 | 61.00 |
| CNN+ transfer | 61.84 | 60.00 | 60.91 | 69.58 | 61.74 | 65.42 |
| ATL | 63.68 | 63.40 | 63.54 | 70.84 | 65.02 | 67.81 |
| ANNSA | 58.82 | 73.34 | 64.18 | - | - | - |
| KESDT | 88.16 | 87.63 | 87.82 | 71.22 | 69.15 | 68.63 |
| **DMFE** | **88.40** | **89.63** | **89.01** | **76.27** | **71.43** | **73.77** |

Table 2: 模型在CADEC、TwiMed数据集上的实验结果

通过实验数据可以看出DMFE考虑到单词拼写、语法不规范问题并且使用Prompt与领域关键词提高模型对领域知识理解之后，在两个数据集上的F1值都取得了比基线模型更好的效果。相较于没有引入预训练模型的工作DMFE有很大的提升，相较于KESDT模型本文除引入了药物不良反应领域知识外还引入字符级信息与AMR特征，解决了社交媒体数据的单词和语法不规范性的问题从而减少数据中的噪声，另外本文还尝试使用Llama2-base(Touvron et al., 2023)通过指令微调训练生成式的检测模型，但由于数据集分布不均衡，导致微调之后很难将社交媒体数据检测为含药物不良反应的数据，并且相较于DMFE占用大量的计算资源，故没有列为合格基线模型作为比较。

### 4.5 多特征提取层的消融实验

为进一步评估模型各维度特征提取的有效性，本文从模型中删除基于Prompt的领域信息、charCNN字符特征提取、AMR-GAT语义特征提取进行了消融实验，实验结果如表3、表4所示。消融之后模型与DMFE相比，在CADEC数据集上的F1值分别下降了1.72%、1.36%和1.04%，在TwiMed数据集上F1值分别下降了4.69%、1.69%和2.02%，充分证明了不同维度特征提取在社交媒体中进行药物不良反应检测起到重要作用。其中消融基于Prompt的领域知识后影响最大，证明Prompt和领域知识可以帮助模型更好地理解药物不良反应，从而提高检测效果。


(c) 2024 中国中文信息学会计算语言学专业委员会

| 模型方法 | P(%) | R(%) | F1(%) |
|---|---|---|---|
| Bert(bert-base-uncased) | 86.76 | 85.43 | 86.09 |
| w/o prompt learning | 87.02 | 87.56 | 87.29 |
| w/o charCNN | 86.32 | 89.02 | 87.65 |
| w/o AMR-GAT | 87.43 | 88.52 | 87.97 |
| DMFE | **88.40** | **89.63** | **89.01** |

Table 3: CADEC数据集上的消融实验

| 模型方法 | P(%) | R(%) | F1(%) |
|---|---|---|---|
| Bert(bert-base-uncased) | 66.71 | 62.38 | 64.47 |
| w/o prompt learning | 69.82 | 68.36 | 69.08 |
| w/o charCNN | 70.98 | 73.22 | 72.08 |
| w/o AMR-GAT | 71.49 | 72.02 | 71.75 |
| DMFE | **76.27** | **71.43** | **73.77** |

Table 4: TwiMed数据集上的消融实验

### 4.6 低资源情况下模型表现结果

社交媒体数据在药物不良反应方向数据集具有稀缺性，所以模型在低资源情况下的表现尤为重要。为验证DMFE在低资源场景下的有效性，分别取训练集的20%、40%、60%、80%数据量在DMFE上进行实验与Bert（Bert-base-uncased）模型进行对比，可以看出DMFE相较于Bert模型在低资源下有更好的性能，实验结果如图5所示。

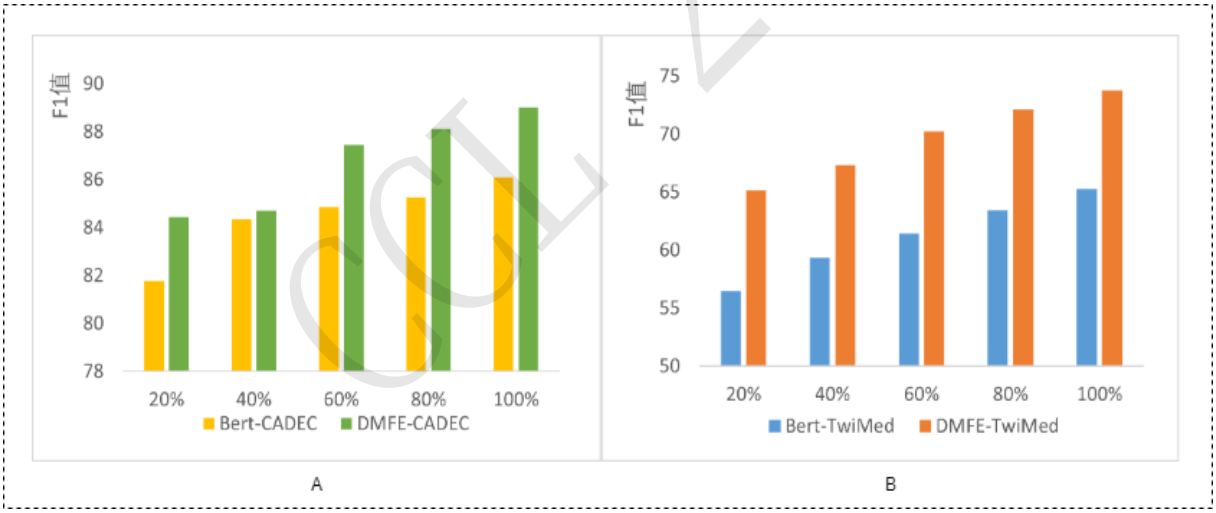



Figure 5: 低资源情况下DMFE与Bert的对比实验

### 4.7 不同Prompt对检测效果的影响

基于提示学习融入领域知识时不同Prompt的选取对模型检测能力影响不同，本文共选取三条不同的Prompt指令进行实验，在CADEC与TwiMed数据集上的实验结果如图6所示，不同Prompt对模型性能有不可忽略的影响。在CADEC数据集上有1.16%的差异，在TwiMed数据集上有1.93%的差异，如何选取更能促进模型理解的Prompt模板对模型检测能力的提升至关重要且值得进一步探讨。


(c) 2024 中国中文信息学会计算语言学专业委员会

| Prompt指令 | CADEC数据集F1值 | TwiMed数据集F1值 |
|---|---|---|
| ADR key words: $K_i$ [SEP] $S$ | 87.85 | 72.17 |
| Is there ADR in the following sentence with the key words: $K_i$ [SEP] $S$ | 88.75 | 74.10 |
| The key words: $K_i$ [SEP]Is there adverse drug reaction in the following sentence? [SEP] $S$ | 89.01 | 73.77 |

Figure 6: 不同Prompt对检测效果的影响

## 5 结束语

为提高在社交媒体中药物不良反应的检测能力，本文充分考虑数据与任务特点使用多特征融合的方式提高检测效果，提出了DMFE模型：利用提示学习的方式融入药物不良反应领域关键词信息，另外采用卷积神经网络charCNN捕获字符级别信息，解决用户单词拼写不准确的问题，结合图注意力网络（GAT）学习句子的抽象语义表示（AMR）提高模型对语义的理解能力，以解决数据语法不规范问题。对DMFE的消融实验证明了各维度特征提取的有效性，此外本文还探讨了低资源环境下模型的有效性，以及不同Prompt模板对实验结果的影响。

在未来，我们对社交媒体中药物不良反应的检测主要集中在两方面：一方面结合生成式模型对药物不良反应的检测模型进行改进，充分发挥生成式模型大量知识信息的优势；另一方面将尝试引入半监督的方式通过其他领域数据进行迁移学习，以解决药物不良反应数据集稀缺的问题。

## 参考文献


Ilseyar Alimova and Valery Solovyev. 2018. Interactive attention network for adverse drug reaction classification. In *Artificial Intelligence and Natural Language: 7th International Conference, AINL 2018, St. Petersburg, Russia, October 17–19, 2018, Proceedings 7*, pages 185–196. Springer.

Nestor Alvaro, Yusuke Miyao, Nigel Collier, et al. 2017. Twimed: Twitter and pubmed comparable corpus of drugs, diseases, symptoms, and their relations. *JMIR public health and surveillance*, 3(2):e6396.

Adrian Benton, Lyle Ungar, Shawndra Hill, Sean Hennessy, Jun Mao, Annie Chung, Charles E Leonard, and John H Holmes. 2011. Identifying potential adverse effects using the web: A new approach to medical hypothesis generation. *Journal of biomedical informatics*, 44(6):989–996.

Yahui Chen. 2015. Convolutional neural network for sentence classification. Master's thesis, University of Waterloo.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

National Center for ADR Monitoring，China. 2023. 国家药品不良反应监测年度报告（2022年）. 中国病毒病杂志, 13:245–251.

Trung Huynh, Yulan He, Alistair Willis, and Stefan Rüger. 2016. Adverse drug reaction classification with deep neural networks. Coling.

Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015. Cadec: A corpus of adverse drug event annotations. *Journal of biomedical informatics*, 55:73–81.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *Eprint Arxiv*.

Michael Kuhn, Monica Campillos, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. 2010. A side effect resource to capture phenotypic effects of drugs. *Molecular systems biology*, 6(1):343.

Zhiheng Li, Zhihao Yang, Ling Luo, Yang Xiang, and Hongfei Lin. 2020. Exploiting adversarial transfer learning for adverse drug reaction detection from texts. *Journal of biomedical informatics*, 106:103431.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Lara Magro, Elena Arzenton, Roberto Leone, Annette Rudolph, and Ugo Moretti. 2021. Identifying and characterizing serious adverse drug reactions associated with drug-drug interactions in a spontaneous reporting database. *Frontiers in Pharmacology*, 11:622862.

Cinzia Marano, Laura Murianni, and Laura Sticchi. 2005. To err is human. building a safer health system. *Epidemiology Biostatistics Public Health*, 2(3-4).

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Patricia Mozzicato. 2009. Meddra: an overview of the medical dictionary for regulatory activities. *Pharmaceutical Medicine*, 23:65–75.

Apurv Patki, Abeed Sarker, Pranoti Pimpalkhute, Azadeh Nikfarjam, Rachel Ginn, Karen O'Connor, Karen Smith, and Graciela Gonzalez. 2014. Mining adverse drug reaction signals from social media: going beyond extraction. *Proceedings of BioLinkSig*, 2014:1–8.

Yunzhi Qiu, Xiaokun Zhang, Weiwei Wang, Tongxuan Zhang, Bo Xu, and Hongfei Lin. 2023. Kesdt: knowledge enhanced shallow and deep transformer for detecting adverse drug reactions. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 601–613. Springer.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *Chinese computational linguistics: 18th China national conference, CCL 2019, Kunming, China, October 18–20, 2019, proceedings 18*, pages 194–206. Springer.

Kanchan Sushim and Gaidhane Abhay. 2023. Social media role and its impact on public health: A narrative review. *Cureus*, 15(1).

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Chuhan Wu, Fangzhao Wu, Junxin Liu, Sixing Wu, Yongfeng Huang, and Xing Xie. 2018. Detecting tweets mentioning drug name and adverse drug reaction with hierarchical tweet representation and multi-head self-attention. In *Proceedings of the 2018 EMNLP workshop SMM4H: the 3rd social media mining for health applications workshop & shared task*, pages 34–37.

Andrew Yates and Nazli Goharian. 2013. Adrtrace: detecting expected and unexpected adverse drug reactions from user reviews on social media sites. In *European Conference on Information Retrieval*, pages 816–819. Springer.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Tongxuan Zhang, Hongfei Lin, Bo Xu, Liang Yang, Jian Wang, and Xiaodong Duan. 2021. Adversarial neural network with sentiment-aware attention for detecting adverse drug reactions. *Journal of Biomedical Informatics*, 123:103896.

CCL 2024

# 面向中文文本的情绪持有者抽取研究

孙亚伟[1]，石宇[1]，韩旭[2,*]
[1]北京邮电大学，计算器学院（国家示范性软件学院），北京，100876
[2]中国科学技术信息研究所，北京，100038
{sunyawei, yus}@bupt.edu.cn，hanx@istic.ac.cn

## 摘要

情绪持有者是文本中带有情绪的主体，对这些情绪持有者的分析对文本情绪理解至关重要。然而，现有研究未充分考虑情绪持有者的共指现象，且由于缺乏面向中文语料的情绪持有者抽取数据，这一研究的发展受到了进一步的限制。本文构建了一个针对中文文本的情绪持有者抽取数据集，有效解决了数据中的共指问题。同时，提出了一种融合语义、情绪和词性特征的模型，实现了高效的情绪持有者抽取与共指消解，且在各项性能指标上超越了基线模型。消融实验进一步证明了模型设计的有效性。[1]

**关键词：** 情绪分析；信息抽取；情绪持有者

# Research on Emotion Holder Extraction for Chinese Texts

**Yawei Sun[1], Yu Shi[1], Xu Han[2,*]**
[1]School of Computer Science (National Pilot Software Engineering School), Beijing University of Posts and Telecommunications, Beijing 100876
[2]Institute of Scientific and Technical Information of China, Beijing 100038
{sunyawei, yus}@bupt.edu.cn, hanx@istic.ac.cn

## Abstract

Emotion holders are the entities in text that exhibit emotions, and their analysis is crucial for understanding emotional expressions in text. However, existing studies have not fully considered the phenomenon of co-reference among emotion holders, and the lack of Chinese corpora has also hindered the development of this research area. This paper constructs a dataset for the extraction of emotion holders from Chinese texts, effectively resolving the co-reference issues within the data. Additionally, a model that integrates semantic, emotional, and part-of-speech features is proposed. It achieves efficient extraction and resolution of emotion holders, surpassing baseline models in various performance metrics. Ablation studies further validate the rationality of the model design.

**Keywords:** Emotion Analysis , Information Extraction , Emotion Holder

## 1 引言

*通讯作者
[1]本研究的数据集及代码可通过如下链接获取：https://github.com/yaweisun/ehe

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

随着互联网的迅猛发展，每天都在产生大量的文本数据。如何从这些数据中提取对人类有益的信息，成为自然语言处理（Natural Language Processing，NLP）领域的一个核心挑战。情绪在研究人类行为和心理健康方面占据了至关重要的地位(Dolan, 2002)。情绪不仅影响个体的心理状态，也会影响其决策过程、人际关系以及身体健康。鉴于文本中包含了丰富的情绪信息，因此在NLP领域，对文本进行情绪分析是一个重要的研究方向。

目前，文本情绪分析主要集中在情绪的分类和识别上，例如识别对话中的情绪(Zhang et al., 2023)以及分析社会热点事件中的情绪表现(Ding et al., 2020)。然而，这些工作仅触及情绪分析的初级阶段。为了更全面地理解文本中蕴含的情绪信息，研究者还进一步挖掘包括情绪线索(Liew et al., 2016)、情绪持有者(Campagnano et al., 2022)、情绪原因(Gui et al., 2016)以及情绪所引发的结果(Sun et al., 2023)在内的复杂情绪信息。图 1展示了这些高级情绪分析任务的实施示例。

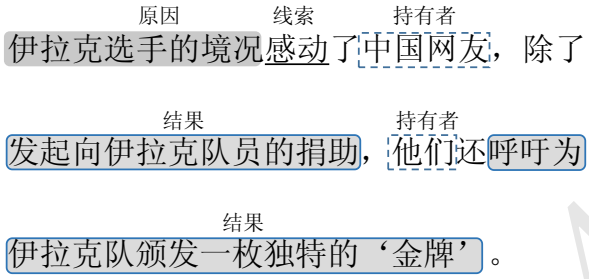



Figure 1: 文本情绪信息抽取示例

在情绪分析中，情绪线索（Emotion Cue）是用于判断文本情绪类别的关键依据。例如，在图 1中，“感动”作为情绪线索，体现了一种积极的情绪。情绪持有者（Emotion Holder）是产生情绪的主体，图中的“中国网友”和“他们”便是示例中的情绪持有者。情绪原因（Emotion Cause）是情绪的触发因素，而情绪所引发的结果通常表现为情绪持有者受情绪影响所采取的行为，即情绪驱动行为（Emotion-Driven Behavior）。这些行为是研究情绪对持有者所造成影响的重要基础，而情绪持有者则是情绪与情绪驱动行为之间的关键连接点。因此，在情绪分析中，对情绪持有者的抽取显得尤为重要。

尽管已经有一系列针对情绪持有者的研究，但这些研究通常未能充分考虑文本中广泛存在的情绪持有者表达方式的共指问题。例如，在图 1所示的示例中，“中国网友”和“他们”这两个短语指代了同一主体，是该主体在文本中的两种不同的表达方式，即它们之间存在共指（Coreference）关系。就语义分析的角度而言，这两种表达方式分别蕴含了不同的语义信息。具体而言，通过“中国网友”这种表达方式，可以分析得到情绪产生的原因；而通过“他们”这种表达方式，可以得到该主体所做出的行为，即情绪引发的结果。然而，现有研究通常只将“中国网友”识别为情绪持有者，而忽略了“他们”。这就使得研究人员难以从语义层面直接识别情绪所引发的后果。在结构更为复杂的文本中，则可能对更全面、深入的情绪信息挖掘造成限制。因而，识别文本中不同表达方式之间的共指关系对于深入、全面的情绪信息分析来说非常有必要，实现这一目标的任务即为共指消解（Coreference Resolution）。此外，中文领域的情绪持有者研究缺乏相应的语料，这进一步限制了这一领域在中文环境中的研究发展。

针对上述问题，本文在中文领域的情绪持有者分析方面进行了深入研究，主要完成了以下工作：

（1）本文以情绪原因抽取（Emotion Cause Extraction，ECE）数据集(Gui et al., 2016)和情绪驱动行为抽取（Emotion-Driven Behavior Extraction，EDBE）数据集(Sun et al., 2023)为基础，构建了一个专门针对中文新闻文本的情绪持有者抽取数据集，该数据集同时实现了情绪持有者的共指消解。

（2）基于新构建的数据集，本文提出了一种新的模型——情绪与词性特征融合的情绪持有者抽取模型（Emotion and Part-of-speech Feature Fusion for Emotion Holder Extraction，EPF-EHE）。EPF-EHE模型结合了文本中的情绪特征、词性特征及语义特征，有效地实现了情绪持有者的抽取和共指消解。并通过对比实验和消融实验验证了模型的有效性和合理性。

第二十三届中国计算语言学大会论文集，第526页-第538页，太原，中国，2024年7月25日至28日。
卷1：主会论文
(c) 2024 中国中文信息学会计算语言学专业委员会

## 2 相关工作

作为文本情绪分析的一个重要研究方向，情绪持有者的抽取在NLP领域得到了众多研究人员的关注。

早期的研究主要依赖基于规则或机器学习的方法来抽取文本中与情绪相关的语义角色。举例来说，Das and Bandyopadhyay (2010b)最初利用规则对孟加拉语博客文本中的情绪持有者进行了标注。随后，他们的工作进一步扩展，探索了使用机器学习方法来识别情绪持有者和情绪原因的可能性(Das and Bandyopadhyay, 2010a; Das and Bandyopadhyay, 2013)。Mohammad et al. (2014)则自动标注了一个与2012年美国大选相关的大型推文数据集，他们结合规则和支持向量机（Support Vector Machine，SVM）对推文进行了情绪分类，并检测了其中的情绪持有者和情绪目标。Ghazi et al. (2015)则基于FrameNet[1]的情绪导向框架构建了一个数据集，该数据集标注了情绪及其持有者等语义角色，并开发了一个条件随机场（Conditional Random Fields，CRF）模型，用于检测文本中的情绪角色信息。

随着深度学习技术的突飞猛进，这一领域涌现出众多优秀的开源数据集。例如，Kim and Klinger (2018)基于古登堡计划（Project Gutenberg）[2]提供的文学作品，手动标注了包含情绪持有者、情绪目标等情绪角色信息的REMAN数据集，并提供了基于CRF和双向长短时记忆网络（Bidirectional Long Short-Term Memory，BiLSTM）的基线模型。Bostan et al. (2020)则创建了一个包含5000个新闻标题的语料库GoodNewsEveryone，这些标题经过众包标注，涵盖了情绪、情绪持有者、情绪线索、情绪目标、情绪原因以及读者对标题情绪的感知，同时提供了一个基于CRF和BiLSTM的基线模型。在上述研究的基础上，Campagnano et al. (2022)提出了一个专注于情绪语义角色标注的评估框架SRL4E，通过采用统一的标注方案来整合多个已标注情绪和语义角色的数据集。

在以BERT(Devlin et al., 2019)为代表的预训练语言模型在NLP各项任务中取得显著成功后，研究人员开始利用这些模型来抽取情绪角色信息。例如，Wegge et al. (2022; 2023)基于BERT等预训练语言模型微调构建了多任务学习模型，以同时抽取各类情绪角色信息。

## 3 情绪持有者抽取数据集

正如第2节中所述，目前已经有国外的研究人员推出了解决情绪持有者抽取问题的数据集。然而，这些数据集并未考虑文本中存在的情绪持有者表达的共指问题。此外，情绪持有者的抽取在中文领域尚缺乏相应数据集的支撑。为了解决这些问题，本文在ECE数据集和EDBE数据集的基础上，构建了针对中文文本的情绪持有者抽取数据集，该数据集同时解决了文本中存在的情绪持有者的共指消解问题。

本节，将详细介绍本文工作所构建的情绪持有者抽取数据集，主要包括数据源的选取与数据预处理、数据标注与数据集的统计。

### 3.1 数据源的选取与数据预处理

本文基于ECE数据集展开研究，保留了数据集中的情绪类别和情绪线索，并对每个情绪线索所对应的情绪持有者进行详细标注。ECE数据集采用了Ekman (1992)提出的情绪分类标准，该标准将基本情绪划分为愤怒（Anger）、厌恶（Disgust）、恐惧（Fear）、喜悦（Happiness）、悲伤（Sadness）和惊讶（Surprise）六类，这一分类方案在W3C情绪标记语言中被称为“Big6”方案[3]。ECE数据集选取2013年至2015年间发布的新浪新闻[4]文本数据进行标注，共获得2105条标注数据。

考虑到ECE数据集的数据量较小，本文进一步选取了THUCNews数据集[5]中“社会”类别下的新闻文本进行扩展。首先，对数据进行清洗，移除如“\u3000”、“ ”等特殊字符。随后，通过识别“。”、“？”、“！”等可视为句子终结的标点符号来对文本进行分句。在得到分句后的文本基础上，本文使用基于BERT的情绪分类模型，该模型是基于ECE数据集训练的，用以对分句后的文本进行情绪分类。若某个句子被判断属于特定情绪类别，便将其选入待标注数据集。

[1] http://framenet.icsi.berkeley.edu/
[2] https://www.gutenberg.org/
[3] https://www.w3.org/TR/emotion-voc/xml#big6
[4] http://news.sina.com.cn/society/
[5] http://thuctc.thunlp.org/


(c) 2024 中国中文信息学会计算语言学专业委员会

### 3.2 数据集的标注

在数据标注阶段，本文使用开源的数据标注工具doccano[6]进行数据标注。

具体而言，本文首先增加了一个用于标注情绪持有者的标签“Holder”。为了保证标注的情绪持有者尽可能包含更加全面的信息，本文所标注的情绪持有者应包含其所有的定语，如“张三的好朋友李四”应标注这一完整表达，而不能只标注“李四”。同时，增加了“情绪-持有者（Emotion-Holder）”和“共指（Coreference）”两个关系标签来解决一段文本存在多个情绪线索的情况下而导致的情绪线索与情绪持有者表达无法准确对应的问题。其中，“Emotion-Holder”为连接情绪线索及与其在语法上存在直接关系的持有者表达的关系，对于每一个情绪线索，虽然它的情绪持有者在文中可能存在多种表达形式，但是本文只将与其在语法上存在直接关系的情绪持有者表达进行标注。“Coreference”则表示共指关系，用于连接情绪持有者的不同表达方式。

为了更加清晰地展示本文的标注方案，本文给出了一个使用doccano对数据进行情绪持有者标注的示例。在该示例中，“感动”一词作为情绪线索代表了一种“喜悦”的情绪，因而其标签为“Happiness”。这一情绪的情绪持有者在文中出现了两处表达方式，分别是“中国网友”和“他们”，因而将这两种表达标注为“Holder”。在关系上，文中出现的第一个“中国网友”是“感动”的宾语，他们之间在语法上存在直接关系，因而将文中的“感动”与“中国网友”之间标注为“Emotion-Holder”关系。文中情绪持有者的两个表达之间则存在共指关系，为了统一数据标注的标准，本文将该情绪持有者的第一个表达之后的其余表达均与第一个表达之间标注为共指关系，即“他们”与“中国网友”之间的关系为“Coreference”。

在进行标注后，本文以情绪线索为基础来形成数据。具体而言，如果同一段文本中有多个情绪线索，则将各情绪线索与其对应的情绪持有者表达单独分离出来构成一条数据。

### 3.3 数据集统计信息

为了理解情绪持有者抽取任务的数据特性，本文对所标注的数据集进行了详尽的统计与分析。这些统计结果不仅揭示了数据集中的基础数量特征，还探究了情绪线索与持有者之间的关系，以及不同情绪类别下的持有者特征分布。

首先，表1展示了情绪持有者抽取数据集的基础统计信息。数据集共包含3971条数据，其中每条数据可能包含一个或多个情绪持有者的表达。从表中可以看出，大多数数据（2008条）包含一个情绪持有者表达，而包含两个或更多持有者的数据则相对较少。此外，平均每条数据包含1.8个情绪持有者的共指表达。

| 项目 | 统计值 |
|---|---|
| 数据量 | 3971 |
| 有1个情绪持有者表达的数据量 | 2008 |
| 有2个情绪持有者表达的数据量 | 1135 |
| 有3个情绪持有者表达的数据量 | 509 |
| 有4个及以上情绪持有者表达的数据量 | 319 |
| 每条数据的平均持有者表达数 | 1.8 |

Table 1: 情绪持有者抽取数据集的统计信息

进一步地，本文对数据集中不同情绪类别的分布情况进行了统计。如表2所示，恐惧和悲伤是数据集中出现频率最高的两种情绪类别，分别占据了总情绪数的25.66%和24.98%。这表明在所研究的文本数据中，这两种情绪是相对较为突出的。相比之下，愤怒、厌恶和喜悦等情绪类别则占据了中等比例，而惊讶情绪的出现频率最低，仅占2.74%。这表明了本文构建的数据集通其他主流数据集一样(Demszky et al., 2020)，也存在数据不平衡的问题。进一步来讲，文章构建的数据集中负面情绪较多，这与其数据来源有密切关系——负面情绪往往更容易引发一些过激行为，从而构成新闻事件。

[6] https://github.com/doccano/doccano


(c) 2024 中国中文信息学会计算语言学专业委员会

| 情绪类别 | 数量 | 百分比 |
|---|---|---|
| 恐惧 | 1019 | 25.66% |
| 悲伤 | 992 | 24.98% |
| 愤怒 | 714 | 17.98% |
| 厌恶 | 487 | 12.27% |
| 喜悦 | 650 | 16.37% |
| 惊讶 | 109 | 2.74% |

Table 2: 情绪持有者抽取数据集中的情绪类别统计

## 4 方法

针对中文情绪持有者抽取问题，本文提出EPF-EHE模型。该模型将情绪持有者的抽取任务看做一个序列标注任务。具体而言，针对文本中的每一个情绪线索，抽取其情绪持有者及情绪持有者的所有共指表达形式。在模型设计上，首先通过语义特征编码层、情绪特征编码层和词性特征编码层来分别获取文本中每个字符的语义特征、情绪特征以及词性特征；然后通过特征融合层将得到的上述三个特征进行融合；接着利用BiLSTM将特征融合层得到的融合特征进行进一步的编码；最后利用CRF对特征进行解码，得到序列标注的结果。

EPF-EHE模型的总体架构如图2所示。接下来，本节将对该模型进行详细介绍。

### 4.1 语义编码层

在自然语言处理任务中，一项基础性的工作就是将文本转化成向量的表示形式，以便于后续对文本数据进行分析与计算。具体而言，对于长度为$n$的输入$X = [x_1, x_2, \cdots, x_n]$，该模块就是将组成$X$的每个$x_i$转变成一个带有上下文语义信息的向量$v_i$，即将X转变为$V = [v_1, v_2, \cdots, v_n]$，其中$i \in [0, n]$。

众所周知，利用最原始的独热编码进行编码后得到的向量不包含语义信息，而利用像Word2Vector这样的静态词向量进行编码，得到的向量虽然包含一定的语义信息，但是这些向量中无法携带上下文信息。因而，本文采用基于Transformer的预训练语言模型作为文本的编码器。为了便于与其他模型进行比较，本文采用经典的基础版$\text{BERT}_{\text{base}}$模型作为编码器。

BERT是一种双向编码表示的预训练语言模型，它在大量文本数据的基础上，基于掩码语言建模和下一句预测这两个无监督的预训练任务来学习语言表示。BERT对文本进行编码，需要在文本的开头加一个句子分类标记（[CLS]），在文本的结尾加一个句子分隔标记（[SEP]）。因而，使用BERT对$X$进行编码，需要首先在$X$的开头加一个句子分类标记$x_{[\text{CLS}]}$，在$X$的结尾加一个句子分隔标记$x_{[\text{SEP}]}$，即$X = [x_{[\text{CLS}]}, x_1, x_2, \cdots, x_n, x_{[\text{SEP}]}]$。相应地，编码得到的$V = [v_{[\text{CLS}]}, v_1, v_2, \ldots, v_n, v_{[\text{SEP}]}]$。对于$\text{BERT}_{\text{base}}$，它输出的向量维度为768，即$V$中每个$v_i$的维度为$1 \times 768$，则$V$的维度为$(n+2) \times 768$。

### 4.2 情绪特征编码层

一段文本中可能存在多个情绪线索，要针对指定的长度为$m$的情绪线索$C_e = [x_{e1}, x_{e2}, \cdots, x_{em}]$来抽取情绪的持有者，那么就需要以某种方式将情绪线索$C_e$的信息传递到本文的情绪持有者抽取模型中。

一种比较简单的方法就是直接将$C_e$与$X$拼接得到一个新的输入$X'$，则$X'$的一种可能取值为$[x_{[\text{CLS}]}, x_{e1}, x_{e2}, \cdots, x_{em}, x_{[\text{SEP}]}, x_1, x_2, \cdots, x_n, x_{[\text{SEP}]}]$，对于图1中的例1而言$X'$对应的文本即为“[CLS]感动[SEP]伊拉克选手的境况感动了中国网友，除了发起向伊拉克队员的捐助，他们还呼吁为伊拉克队颁发一枚独特的‘金牌’。[SEP]”。然而，这就存在这样一个问题：一段文本中可能会有多个相同的情绪线索，而这些情绪所对应的持有者是不同的。例如，对于这样的一段文本：“警察制服了情绪激动的绑匪，围观群众激动地鼓起掌来”，其中“激动”出现了两次，然而这两个“激动”所表达的情绪其实是不同的。第一个“激动”表达的是一种“愤怒”的情绪，这一情绪的持有者是“情绪激动的绑匪”。而第二个“激动”则表达了一种“快乐”的情绪，这一情绪的持有者则是“围观群众”。显然，这种情况下，如果直接采用拼接的方法构成新的输入$X'$，则容易对模型造成困惑，最终无法得到正确的结果。


(c) 2024 中国中文信息学会计算语言学专业委员会

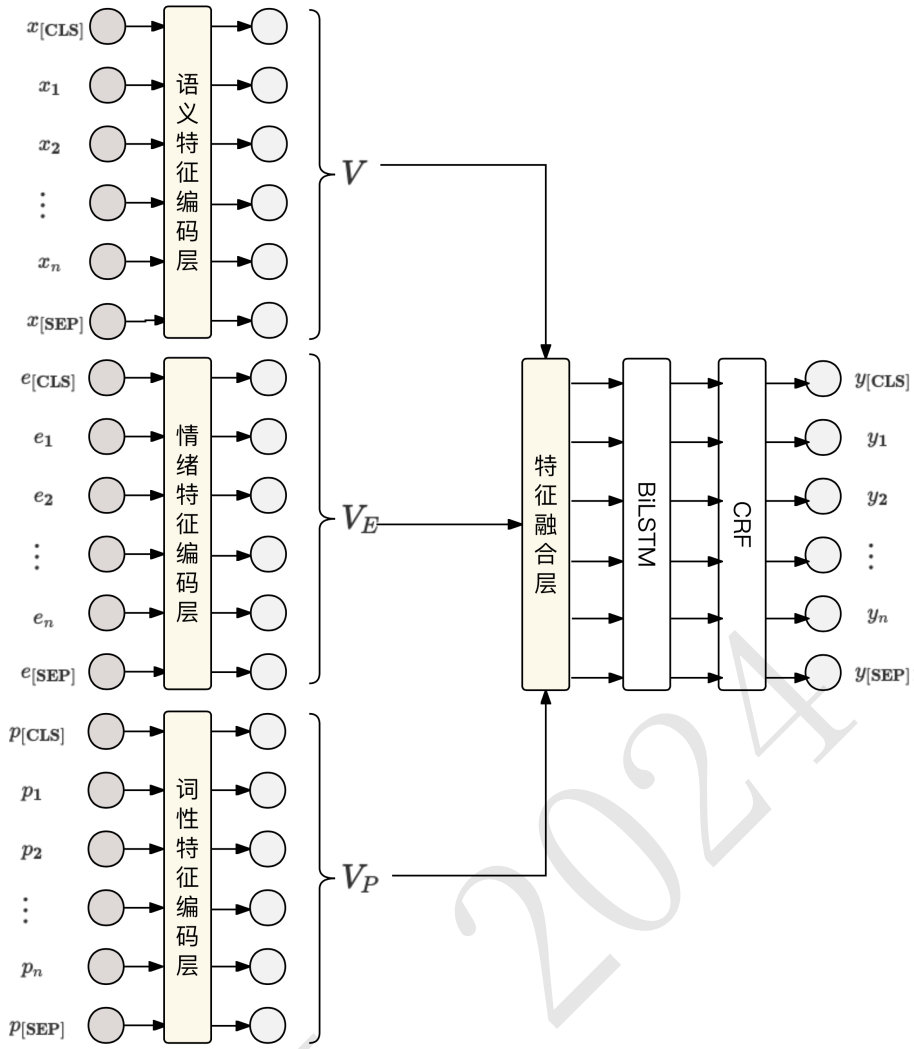



Figure 2: EPF-EHE模型框架图

为了解决这一问题，本文在情绪持有者抽取模型中增加了对情绪线索位置的编码层，即情绪特征编码层。具体而言，就是将指定的情绪线索在原始文本中对应的位置标为1，其他位置则标为0，进而得到情绪编码序列$E = \{e_{[\text{SEP}]}, e_1, e_2, \cdots, e_n, e_{[\text{SEP}]}\}$，其中$e_i \in \{0, 1\}$。由于[CLS]与[SEP]肯定不会属于任意一个情绪线索的组成部分，因而$e_{[\text{SEP}]}$与$e_{[\text{CLS}]}$默认为0。对于情绪线索$C_e = \{x_{e1}, x_{e2}, \cdots, x_{em}\}$，$E$中的元素$e_i$计算方法如式（1）所示。

$$e_i = \begin{cases} 1, & i \in [e1, em] \\ 0, & \text{其他} \end{cases} \tag{1}$$

为了将情绪特征与文本$X$经由语义编码层编码得到的语义特征$V$相融合，本文通过增加一个嵌入层$Embding_E$将$E$编码成与$V$形状相同的向量$V_E$，如式（2）所示。

$$V_E = Embding_E(E) \tag{2}$$

嵌入层本质上是一种词典，其中保存了它所要编码的状态所对应的向量。由于$p_{ei}$只有0和1两种状态，而且经过嵌入层编码$Embding_E$得到的情绪特征向量$V_E$需要与语义特征向量$V$一致，因而，本文对情绪特征进行编码的嵌入层$Embding_E$的形状为$2 \times 768$。根据$V_E$的定义可知，其长度为$n+2$，而$Embding_E$的形状为$2 \times 768$，则$V_E$的形状为$(n+2) \times 768$。

嵌入层的参数是可训练的，因而，在训练前需要先对$Embding_E$进行初始化。在训练过程中，模型可以动态更新$Embding_E$的参数，最终达到对情绪特征进行编码的效果。


(c) 2024 中国中文信息学会计算语言学专业委员会

### 4.3 词性特征编码层

情绪持有者，从定义上来讲就是名词性的词语或短语。从本文的研究目标来看，本文抽取得到的情绪持有者可能是包含了完整定语的人物、机构，也可能是代词。而情绪持有者在文本的语言表达上定语则可能是形容词等其他词性的词。如果能够将文本中的命名实体和词性信息作为特征融合到情绪持有者抽取模型中，很可能将有助于模型效果的提升。

这涉及自然语言处理领域的命名实体识别和词性标注这两个基础性任务。目前，已经有很多比较成熟的工具可以帮助用户来处理这方面的工作，像LAC(Jiao et al., 2018)[7]、LTP(Che et al., 2021)[8]、HarvestText[9]以及HanLP(He and Choi, 2021)[10]等，都是在中文自然语言处理领域非常优秀的开源工具。上述提到的开源工具，多数都是将命名实体识别与词性标注看做两个独立的任务，如果采用这样的工具，则需要对文本中的命名实体和词性进行分别处理。然而，从命名实体的定义上来讲，它就是名词性的短语，将它与词性标注分开来处理在本文任务中显然是没有必要的。百度公司开发的LAC则不存在这样的问题。LAC将命名实体识别与词性标注整合为一个任务，其将命名实体类别也视为词性，采用统一的接口来实现。LAC的包含命名实体类别在内的词性类别标签如表3所示。

| 序号 | 词性标签 | 含义 | 序号 | 词性标签 | 含义 | 序号 | 词性标签 | 含义 |
|---|---|---|---|---|---|---|---|---|
| 0 | n | 普通名词 | 9 | ORG | 机构名 | 18 | an | 名形词 |
| 1 | nz | 其他专名 | 10 | nw | 作品名 | 19 | r | 代词 |
| 2 | a | 形容词 | 11 | vn | 名动词 | 20 | d | 副词 |
| 3 | m | 数量词 | 12 | ad | 副形词 | 21 | p | 介词 |
| 4 | c | 连词 | 13 | q | 量词 | 22 | w | 标点符号 |
| 5 | PER | 人名 | 14 | u | 助词 | 23 | TIME | 时间 |
| 6 | f | 方位名词 | 15 | LOC | 地名 | 24 | o | 其他符号 |
| 7 | v | 普通动词 | 16 | s | 处所名词 | | | |
| 8 | xc | 其他虚词 | 17 | vd | 动副词 | | | |

Table 3: LAC的词性类别标签信息表

本文利用LAC对文本进行词性标注，得到的是该文本分词后的词语（或短语）列表以及相应的词性标签列表。要想完成对词性信息的编码，需要先对每个词性类别进行编号。本文对每一个词性类别都进行了编号，如表3的“序号”列所示。考虑到EPF-EHE模型在语义编码层对数据进行处理时，需要分别在文本的开头和结尾添加[CLS]和[SEP]这两个符号，同时，对文本进行补齐（Padding），这三种符号不属于原始文本的组成部分，也就不具备任何的词性。为了处理这种情况，本文在表3中增加了标签“o”，代表上述提到的这三种情况。

为了直观地展示使用LAC对文本进行词性标注及相应的编号结果，本文同样以图1中的示例为例。在对中文文本中的词汇进行词性标注时，通常需要先对文本进行分词，然后再以词为单位进行词性标注。然而，BERT对中文文本进行处理并不是基于词汇这一级别的，而是基于字符级别。也就是说，BERT无法直接处理“伊拉克”这样的词汇，而需要先将“伊拉克”分隔为“伊”、“拉”和“克”三个字，然后分别处理这三个字。在示例中，“伊拉克”的词性标签为“LOC”，则本文将它分隔为单个字符后，使组成它的“伊”、“拉”、“克”的词性标签与它本身保持一致。对示例进行词性标注得到的其他词汇，本文均采用相同的处理方法，得到的结果如表4 所示。

表4中“编号”列可组成文本$X$的词性序列$P = [p_{[\mathrm{CLS}]}, p_1, p_2, \cdots, p_n, p_{[\mathrm{SEP}]}]$。与情绪特征编码层相同，想要完成对词性特征的编码，同样需要一个嵌入层$Embding_P$。表3中一共列举了25中词性类型，那么$Embding_P$的形状相应地应该为$25 \times 768$。

通过使用嵌入层$Embding_P$对词性词性序列POS进行编码，可以得到$X$的词性特征编码结果$V_P$，如式（3）所示。POS的长度显然为$n+2$，而$Embding_P$的形状为$25 \times 768$，则$V_P$的形状

[7] https://github.com/baidu/lac
[8] https://github.com/HIT-SCIR/ltp
[9] https://github.com/blmoistawinde/HarvestText
[10] https://github.com/hankcs/HanLP


(c) 2024 中国中文信息学会计算语言学专业委员会

| 词汇 | 词性标签 | 词性编号 | 词汇 | 词性标签 | 词性编号 | 词汇 | 词性标签 | 词性编号 |
|---|---|---|---|---|---|---|---|---|
| [CLS] | o | 24 | 了 | p | 21 | 为 | v | 7 |
| 伊 | LOC | 15 | 发 | v | 7 | 伊 | ORG | 9 |
| 拉 | LOC | 15 | 起 | v | 7 | 拉 | ORG | 9 |
| 克 | LOC | 15 | 向 | p | 21 | 克 | ORG | 9 |
| 选 | n | 0 | 伊 | LOC | 15 | 队 | ORG | 9 |
| 手 | n | 0 | 拉 | LOC | 15 | 颁 | v | 7 |
| 的 | u | 14 | 克 | LOC | 15 | 发 | v | 7 |
| 境 | n | 0 | 队 | n | 0 | 一 | m | 3 |
| 况 | n | 0 | 员 | n | 0 | 枚 | m | 3 |
| 感 | v | 7 | 的 | u | 14 | 独 | a | 2 |
| 动 | v | 7 | 捐 | vn | 11 | 特 | a | 2 |
| 了 | u | 14 | 助 | vn | 11 | 的 | u | 14 |
| 中 | LOC | 15 | , | w | 22 | ‘ | w | 22 |
| 国 | LOC | 15 | 他 | r | 19 | 金 | n | 0 |
| 网 | n | 0 | 们 | r | 19 | 牌 | n | 0 |
| 友 | n | 0 | 还 | d | 20 | ’ | w | 22 |
| , | w | 22 | 呼 | v | 7 | 。 | w | 22 |
| 除 | p | 21 | 吁 | v | 7 | [SEP] | o | 24 |

Table 4: 文本的字符级词性标注示例

为$(n+2)\times 768$。

$$V_P = Embding_P(P) \tag{3}$$

与$Embding_E$相同，本文在训练前需要先对$Embding_P$进行初始化。在训练过程中，模型同样的可以动态更新$Embding_P$的参数，最终达到对词性特征进行编码的效果。

### 4.4 特征融合层

对长度为$n$的文本$X$，分别经过语义编码层、情绪特征编码层和词性特征编码层的编码之后，得到$V$、$V_E$和$V_P$这三个形状均为$(n+2)\times 768$的特征向量。参考BERT在嵌入层对符号嵌入、位置嵌入和段落嵌入这三种嵌入的处理方法，将得到的三个特征向量逐元素相加，从而得到融合后的特征向量$V_c$，如式（4）所示。

$$V_c = V + V_E + V_P \tag{4}$$

### 4.5 基于融合特征向量的情绪持有者抽取

在经过语义编码层、情绪特征编码层和词性特征编码层的编码，并通过特征融合层对三种编码后的特征向量进行融合后，得到了融合了上下文语义信息、情绪信息、词性信息的特征向量$V_c$。接下来，本文的主要任务就是利用这样的特征向量得到最终的抽取结果。

本文将情绪持有者的抽取任务看做一个序列标注的任务来完成。即对于输入$X = [x_{[\text{CLS}]}, x_1, x_2, \cdots, x_n, x_{[\text{SEP}]}]$，需要得到序列标签$Y = [y_{[\text{CLS}]}, y_1, y_2, \cdots, y_n, y_{[\text{SEP}]}]$。这就涉及对组成$X$的字符$x_i$进行标注的问题，本文采用经典的“BIO”标注方案来完成标注。具体而言，就是将X中组成情绪持有者的字符标注成“B”或者“I”，将其他字符标注为“O”。对于组成情绪持有者的字符，如果它是相应的情绪持有者的第一个字符，则将其标注为“B”，否则标注为“I”。

为了更加清晰地阐述本文的标注方案，本文给出了针对图1所示的示例中根据“感动”这一情绪线索所抽取到的情绪持有者的标注结果，如表5所示。

在得到融合特征向量$V_c$以后，是难以直接得到上述这种标注结果的。为了进一步捕获$V_c$的上下文语义信息，本文在特征融合层之后增加了一个BiLSTM层来对特征向量进行进一步的编码。BiLSTM能够从两个方向捕获序列的上下文依赖关系，能够进一步增强每一个字符所对应特征向量的语义表示能力。在经过BiLSTM进行进一步的编码后，得到$V_c$的隐藏向量表示$H$，如式（5）所示。

$$H = \text{BiLSTM}(V_c) \tag{5}$$


(c) 2024 中国中文信息学会计算语言学专业委员会

| 字符 | 标签 | 字符 | 标签 | 字符 | 标签 | 字符 | 标签 | 字符 | 标签 | 字符 | 标签 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| [CLS] | O | 感 | O | 了 | O | 的 | O | 为 | O | 独 | O |
| 伊 | O | 动 | O | 发 | O | 捐 | O | 伊 | O | 特 | O |
| 拉 | O | 了 | O | 起 | O | 助 | O | 拉 | O | 的 | O |
| 克 | O | 中 | B | 向 | O | ， | O | 克 | O | ‘ | O |
| 选 | O | 国 | I | 伊 | O | 他 | B | 队 | O | 金 | O |
| 手 | O | 网 | I | 拉 | O | 们 | I | 颁 | O | 牌 | O |
| 的 | O | 友 | I | 克 | O | 还 | O | 发 | O | ’ | O |
| 境 | O | ， | O | 队 | O | 呼 | O | 一 | O | 。 | O |
| 况 | O | 除 | O | 员 | O | 吁 | O | 枚 | O | [SEP] | O |

Table 5: 情绪持有者抽取的标注示例

接下来，需要对经过上述的一系列编码得到向量表示进行解码。本节中，采用CRF来对$H$进行解码。不同于简单地为每个词独立地选择标签，CRF考虑了标签序列的约束，优化整个序列的标签选择。这意味着模型在为一个词选择标签时，会考虑到整个句子中所有词的标签，确保最终的标签序列在全局上是最优的。经过运用CRF进行解码，最终得到了对输入序列$X$的预测，如式（6）所示。

$$\hat{Y} = \mathrm{CRF}(H) \tag{6}$$

式中，$\hat{Y}$则表示模型对输入序列$X$的预测，其中$\hat{Y} = \{\hat{y}_{[\mathrm{CLS}]}, \hat{y}_1, \hat{y}_2, \cdots, \hat{y}_n, \hat{y}_{[\mathrm{SEP}]}\}$。

通常情况下，在模型训练之初，预测值$\hat{Y}$与标签$Y$之间会存在较大的差异。而本文训练模型的目的就是减少$\hat{Y}$与Y之间的差异。

## 5 实验

本节，将介绍本章工作所构建的数据集，同时将介绍文章工作的实验设置、实验结果及分析等内容。

### 5.1 实验设置

本章在语义编码层采用BERT针对汉语的基础版本，即“bert-base-chinese”。所有实验均使用Tesla V100-SXM2-16GB GPU进行加速。模型采用PaddlePaddle框架来实现。在训练过程中，采用交叉熵损失作为损失函数，使用AdanW作为优化器来最小化损失。在训练过程中，取验证集上$F_1$值最高的模型参数作为最佳模型参数。

在超参数方面，BERT的学习率设置为$1 \times 10^{-5}$，其他部分的学习率则设置为$1 \times 10^{-3}$。实验中未设置最大训练步数，而是采用早停策略来控制训练过程。具体而言，每训练20步即使用验证集进行一次验证，若模型在连续20次验证中未能取得更好的效果，则终止训练。

### 5.2 评价指标

本章采用精确率（Precision，$P$）、召回率（Recall，$R$）和$F_1$值作为实验的评价指标。对于模型预测的一种情绪持有者表达，如果能够与人工标注的情绪持有者表达完全匹配上，则为真阳性（True Positive，$TP$）；如果模型预测的情绪持有者表达无法与人工标注的情绪持有者表达匹配上，则为假阳性（False Positive，$FP$）；如果人工标注的情绪持有者表达没有被模型预测到，则为假阴性（False Negative，$FN$）。则精确率、召回率和$F_1$值的计算公式如下：

$$P = \frac{TP}{TP + FP} \tag{7}$$

$$R = \frac{TP}{TP + FN} \tag{8}$$

$$F_1 = \frac{2 \cdot P \cdot R}{P + R} \tag{9}$$

第二十三届中国计算语言学大会论文集，第526页-第538页，太原，中国，2024年7月25日至28日。
卷1：主会论文
(c) 2024 中国中文信息学会计算语言学专业委员会

### 5.3 基线模型

首先，本章选取如下模型作为基线模型：

**BERT**：该模型是以BERT为基础进行文本序列标注任务的最基础的模型。该模型以BERT作为编码器，以Softmax作为解码器。是HuggingFace[11]与PaddlePaddle[12]官方提供的基线模型。

**BERT-CRF**：该模型以BERT编码器，以CRF作为解码器，被广泛应用于文本序列标注任务中(Souza et al., 2019)。

**BERT-BiLSTM-CRF**：该模型是各类文本序列标注任务中的经典模型。该模型以BERT和BiLSTM作为编码器，以CRF作为解码器，在文本的各类序列标注任务中均取得了较好的效果，并被众多研究作为基线模型(Bostan et al., 2020)。

**SRL4E**：该模型是Campagnano et al. (2022)在其情绪角色标注统一评估框架中提出的模型。对于文本$X$和情绪线索$E$，SRL4E构造了新的文本序列$X' = [\text{CLS}]X[\text{SEP}]E[\text{SEP}]$，将$X'$利用BERT进行编码后，再进一步利用两层的BiLSTM进行语义编码，最终利用Softmax进行解码。

**Code4Struct**：该方法是Wang et al. (2023)针对事件论元抽取任务所提出的基于大预言模型（Large Language Model，LLM）的方法。该方法利用LLM强大的文本理解与推理能力，将自然语言处理任务转化为结构化的代码，进而完成将文本转化为结构化数据的任务。

**ChatGPT 3.5**[13]：该模型是OpenAI公司推出的生成式大语言模型。本文通过对话形式，直接就情绪持有者抽取任务对ChatGPT进行询问，从而得到答案，并通过文本匹配的方式进行各项指标的统计与计算。

### 5.4 结果与分析

（1）情绪持有者抽取的实验结果与分析

表6详细展示了在情绪持有者抽取任务中，EPE-EHE与其他基线模型的比较结果。从表中可以看出，EPE-EHE在精确率、召回率和$F_1$分数这三个关键指标上均取得了最佳表现。

具体来说，EPE-EHE的精确率达到了82.39%，这意味着在模型预测为情绪持有者的样本中，有82.39%是真正的情绪持有者。这一结果明显高于其他基线模型。其中，BERT作为表现次优的模型，其精确率为81.05%，相较于EPE-EHE低了1.34个百分点。

在召回率方面，EPE-EHE同样表现出色，达到了87.66%。这意味着在实际为情绪持有者的样本中，有87.66%被EPE-EHE正确识别。在基线模型中，经典的BERT-BiLSTM-CRF模型的召回率反而更接近EPE-EHE，这主要是因为该模型未考虑情绪线索，这也导致其在精确率上表现最差。

需要注意的是，基于LLM的Code4Struct和ChatGPT 3.5在各项指标上的效果均不理想。虽然主流的LLM模型在文本的语义理解及推理上取得了令人惊叹的效果，然而，在大规模的数据处理上，特别是信息抽取的相关任务中，还存在一定的缺陷。在本研究中，通过试验发现，在进行多轮回话的过程中，LLM比较严重地依赖提示信息，回答的结果不够稳定，在一些问题上会自行扩展一些额外的信息，容易出现答非所问的情况，这些都导致了其在任务中表现不佳的情况。

综合精确率和召回率，EPE-EHE的$F_1$分数达到了84.94%，这一综合指标同样高于其他所有基线模型。在基线模型中，BERT和BERT-BiLSTM-CRF分别在精确率和召回率上达到了接近EPE-EHE的水平，然而他们分别在召回率和精确率上取得了最差的效果，因而综合表现与EPE-EHE之间存在差距。而SRL4E的各项指标表现得中规中矩，Code4Struct与ChatGPT 3.5则在各项指标中均效果不佳。因而EPE-EHE在综合性能上在基线模型中取得了相对领先的效果。

（2）消融实验

消融实验的目的是评估模型各个组成部分的贡献。通过依次移除模型的某一部分（例如，特定的特征或层），可以观察到模型性能的变化，从而理解该部分对模型整体性能的影响。在

[11] https://github.com/huggingface/transformers
[12] https://github.com/PaddlePaddle/PaddleNLP
[13] https://chatgpt.com


(c) 2024 中国中文信息学会计算语言学专业委员会

| 模型 | 精确率 | 召回率 | $F_1$ |
|---|---|---|---|
| BERT | 81.05% | 78.67% | 79.84% |
| BERT-CRF | 74.42% | 81.84% | 77.95% |
| BERT-BiLSTM-CRF | 66.12% | 86.52% | 74.96% |
| SRL4E | 74.56% | 81.84% | 78.03% |
| Code4Struct | 22.26% | 24.70% | 23.42% |
| ChatGPT 3.5 | 13.98% | 19.63% | 16.33% |
| EPF-EHE | **82.39%** | **87.66%** | **84.94%** |

Table 6: 情绪持有者抽取的实验结果

情绪直接持有者抽取任务和情绪持有者抽取任务中，EPF-EHE模型的消融实验结果展示了移除不同组件后模型性能的变化。

| $V_E$ | $V_p$ | BiLSTM | CRF | 精确率 | 召回率 | $F_1$ |
|---|---|---|---|---|---|---|
| + | + | + | + | **82.39%** | **87.66%** | **84.94%** |
| - | + | + | + | 78.86% | 83.79% | 81.25% |
| + | - | + | + | 81.03% | 82.62% | 81.82% |
| + | + | - | + | 70.80% | 78.13% | 74.28% |
| + | + | + | - | 70.85% | 85.94% | 77.67% |

Table 7: 情绪持有者抽取消融实验结果

表7展示了消融实验结果。消融实验结果表明，情绪特征编码、词性特征编码、BiLSTM层和CRF层都对EPF-EHE模型在情绪持有者抽取任务中的性能起到了积极作用。特别是BiLSTM层的移除导致了最为显著的性能下降，这说明上下文信息对于情绪持有者的抽取至关重要。情绪特征编码和词性特征编码的贡献也不容忽视，它们通过提供关于情绪线索位置和词性的信息，帮助模型更精准地定位和理解情绪持有者。尽管CRF层的移除对性能的影响相对较小，但它在优化整个序列的标签选择中仍然发挥着重要作用，特别是在保证标签序列全局一致性方面。

这些结果证明了模型不同组件在处理复杂的情绪持有者抽取任务时的综合作用。情绪特征编码和词性特征编码通过提供关键的文本信息，增强了模型对情绪持有者的识别能力；BiLSTM层通过捕获句子的上下文依赖关系，增强了模型对情绪持有者与情绪线索之间关系的理解；CRF层则在序列标注任务中优化了标签序列的全局一致性。

## 6 总结

本文探讨了文本情绪分析中一个重要但经常被忽视的组成部分——情绪持有者的识别和抽取。情绪持有者是指在文本中显示出特定情绪的实体。准确识别这些实体对于深入理解文本中的情绪信息至关重要。

首先，本文分析了现有研究中的主要挑战，特别是情绪持有者共指现象的处理问题以及针对中文文本的资源匮乏问题。为了应对这些挑战，本文构建了一个专门针对中文文本的情绪持有者抽取数据集。该数据集不仅丰富了中文情绪分析的资源，还通过精确的标注解决了共指现象的问题。

其次，本文提出了一种新的情绪持有者抽取模型，该模型综合了语义特征、情绪特征和词性特征，通过这种融合方法显著提高了情绪持有者抽取的准确性。模型采用BERT作为基础编码器，利用BiLSTM来进一步捕获情绪表达的上下文关系，最后通过CRF层优化标注序列，提高了整体的标注效果。

在实验部分，通过与多个基线模型的对比，验证了所提模型在多个性能指标上的优越性。此外，本文通过消融实验，进一步证明了模型中各个组件的有效性和必要性。

下一步，将以情绪持有者抽取为基础，对文本的情绪分析进行更深层次的研究。

第二十三届中国计算语言学大会论文集，第526页-第538页，太原，中国，2024年7月25日至28日。
卷1：主会论文
(c) 2024 中国中文信息学会计算语言学专业委员会

## 参考文献

Laura Bostan, Evgeny Kim, and Roman Klinger. 2020. GoodNewsEveryone: A Corpus of News Headlines Annotated with Emotions, Semantic Roles, and Reader Perception. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1554–1566.

Cesare Campagnano, Simone Conia, and Roberto Navigli. 2022. SRL4E – Semantic Role Labeling for Emotions: A Unified Evaluation Framework. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4586–4601, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wanxiang Che, Yunlong Feng, Libo Qin, and Ting Liu. 2021. N-LTP: An Open-source Neural Language Technology Platform for Chinese. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 42–49, Stroudsburg, PA, USA. Association for Computational Linguistics.

Dipankar Das and Sivaji Bandyopadhyay. 2010a. Finding Emotion Holder from Bengali Blog Texts-An Unsupervised Syntactic Approach. In *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*, pages 621–628.

Dipankar Das and Sivaji Bandyopadhyay. 2010b. Labeling Emotion in Bengali Blog Corpus – A Fine Grained Tagging at Sentence Level. In Sarmad Hussain, Virach Sornlertlamvanich, and Hammam Riza, editors, *Proceedings of the Eighth Workshop on Asian Language Resouces*, pages 621–628, Beijing. Coling 2010 Organizing Committee.

Dipankar Das and Sivaji Bandyopadhyay. 2013. Emotion Co-referencing-Emotional Expression, Holder, and Topic. *International Journal of Computational Linguistics \& Chinese Language Processing*, 18(1):79–98.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A Dataset of Fine-Grained Emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Keyang Ding, Jing Li, and Yuji Zhang. 2020. Hashtags, Emotions, and Comments: A Large-Scale Dataset to Understand Fine-Grained Social Emotions to Online Topics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1376–1382, Stroudsburg, PA, USA. Association for Computational Linguistics.

R. J. Dolan. 2002. Emotion, Cognition, and Behavior. *Science*, 298(5596):1191–1194, 11.

Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200, 5.

Diman Ghazi, Diana Inkpen, and Stan Szpakowicz. 2015. Detecting Emotion Stimuli in Emotion-Bearing Sentences. In *Computational Linguistics and Intelligent Text Processing*, volume 9042, pages 152–165. Springer International Publishing, Cham.

Lin Gui, Dongyin Wu, Ruifeng Xu, Qin Lu, and Yu Zhou. 2016. Event-Driven Emotion Cause Extraction with Corpus Construction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1639–1649, Stroudsburg, PA, USA. Association for Computational Linguistics.

Han He and Jinho D. Choi. 2021. The Stem Cell Hypothesis: Dilemma behind Multi-Task Learning with Transformer Encoders. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5555–5577, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zhenyu Jiao, Shuqi Sun, and Ke Sun. 2018. Chinese Lexical Analysis with Deep Bi-GRU-CRF Network. *arXiv preprint arXiv:1807.01882*, 7.


(c) 2024 中国中文信息学会计算语言学专业委员会

Evgeny Kim and Roman Klinger. 2018. Who Feels What and Why? Annotation of a Literature Corpus with Semantic Roles of Emotions Title and Abstract in German. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1345–1359, Santa Fe, New Mexico, USA.

Jasy Liew, Suet Yan, Howard R Turtle, and Elizabeth D Liddy. 2016. EmoTweet-28: A Fine-Grained Emotion Corpus for Sentiment Analysis. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1149–1156.

Saif Mohammad, Xiaodan Zhu, and Joel Martin. 2014. Semantic Role Labeling of Emotions in Tweets. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 32–41, Stroudsburg, PA, USA. Association for Computational Linguistics.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2019. Portuguese Named Entity Recognition using BERT-CRF. *arXiv preprint arXiv:1909.10649*, 9.

Yawei Sun, Saike He, Xu Han, and Ruihua Zhang. 2023. A New Model for Emotion-Driven Behavior Extraction from Text. *Applied Sciences*, 13(15):8700, 7.

Xingyao Wang, Sha Li, and Heng Ji. 2023. Code4Struct: Code Generation for Few-Shot Event Structure Prediction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3640–3663, Stroudsburg, PA, USA. Association for Computational Linguistics.

Maximilian Wegge and Roman Klinger. 2023. Automatic Emotion Experiencer Recognition. In *Proceedings of the 3rd Workshop on Computational Linguistics for the Political and Social Sciences*, pages 1–7.

Maximilian Wegge, Enrica Troiano, Laura Ana Maria Oberlaender, and Roman Klinger. 2022. Experiencer-Specific Emotion and Appraisal Prediction. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 25–32, Stroudsburg, PA, USA. Association for Computational Linguistics.

Duzhen Zhang, Feilong Chen, and Xiuyi Chen. 2023. DualGATs: Dual Graph Attention Networks for Emotion Recognition in Conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7395–7408, Stroudsburg, PA, USA. Association for Computational Linguistics.

CCL 2024

# AutoRG：一种大小模型协同的自动报告生成框架

张京[1,2], 舒江明[1,2], 张宇翔[1,2], 吴斌[3], 王巍[3], 于剑[1,2], 桑基韬[1,2]*
[1]北京交通大学计算机科学与技术学院
[2]交通数据分析与挖掘北京市重点实验室
[3]深圳素问数据智能科技有限公司
{zhangj_, 20281016, yuxiangzhang, jianyu, jtsang}@bjtu.edu.cn
{wubin, wangwei}@suwen.ai

## 摘要

自动报告生成技术在提高工作效率和节约人力资源方面具有显著潜力。大语言模型的出现使得报告流畅度与可解释性得到提升。然而，现有工作仍依赖人工，缺乏灵活性和丰富度。同时，小模型错误或冗余的输出与大模型自身的随机性会导致报告质量不稳定。本文提出大小模型协同的自动报告生成框架AutoRG，通过大模型的工具理解与规划能力减少人工干预，提升报告丰富度，并通过信息修正与报告迭代机制提高报告的稳定性。本文以自动专利报告生成为场景，从多个维度对AutoRG进行全面测试。结果表明，该框架在提高报告生成的丰富度和质量稳定性方面具有显著优势。

**关键词**： 自动报告生成； 大语言模型； 智能体； 大小模型协同

# AutoRG: An automatic report generation framework for Large and small model collaboration

**Jing Zhang[1,2], Jiangming Shu[1,2], Yuxiang Zhang[1,2], Bin Wu[3], Wei Wang[3], Jian Yu[1,2], Jitao Sang[1,2]**
[1]School of Computer Science and Technology, Beijing Jiaotong University, Beijing, China
[2]Lab of Traffic Dala Analysis and Mining, Beijing, China
[3]Suwen Data Intelligent Technology Co., LTD., Shenzhen, China
{zhangj_, 20281016, yuxiangzhang, jianyu, jtsang}@bjtu.edu.cn
{wubin, wangwei}@suwen.ai

## Abstract

Automatic report generation technology has demonstrated enormous potential in saving human resources and enhancing work efficiency. With the introduction of large-scale artificial intelligence models, the fluency and interpretability of generated reports have been significantly improved. However, current automatic report generation still requires manual completion, lacking flexibility and richness. Moreover, errors or redundancies in the outputs of small models, along with the inherent randomness of large models, can result in the lack of stability in report content. This paper introduces AutoRG, an integrated framework designed for automated report generatio, to reduce human intervention by leveraging the understanding and planning capabilities of large models, consequently enriching the depth of generated reports. Additionally, it incorporates an information correction and content iteration mechanism to diminish reliance on the outputs of small models and mitigate the impact of randomness inherent in large models. To evaluate the efficacy of the AutoRG framework comprehensively,

*通讯作者
©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版
**基金项目**：中央高校基本科研业务费专项资金(62172094)；国家自然科学基金(62172094)

we conducted experiments focusing on the automated generation of patent technology reports. Our evaluation approach across various dimensions, ensuring a thorough examination of AutoRG's performance. The results indicate significant advantages of the framework in improving the richness and stability of report generation.

**Keywords:** Automatic report generation , Large language model , Agent , Collaboration between large and small models

## 1 引言

自动报告生成是通过计算机算法自动提取原始数据或文本中的关键信息，按照既定格式和结构生成具有特定目的和内容的报告的过程。该技术已广泛应用于多个领域，例如医学影像报告(Nakaura et al., 2024)、社会金融报告(Ren et al., 2021)、企业财务报告(**?**)、信息检索报告(Babour and Khan, 2021)等。自动报告生成的优势在于，它能够减轻人们在报告编写任务上的负担，节省阅读文献、分析数据和重复文本编写工作所需的大量时间和精力。此外，自动生成的报告具有一定的参考价值，能够为行业专家提供撰写报告的参考，甚至作为内容补充。

传统自动报告生成方法主要分为抽取式和生成式两大类。抽取式方法主要通过自然语言处理的小模型，通过规则筛选(Gong et al., 2017)、主题聚类(Wang et al., 2019)、数据统计(Babour and Khan, 2021)等技术手段直接从原始数据中提取相关信息并填充到预设的模板中。这种方法的优势在于能够迅速获取基础信息，但其灵活性较差，难以适应多变的报告需求，同时也缺乏对内容进行深入分析的能力。另一方面，生成式方法通过端到端训练的模式，如编码器-解码器模型，试图模仿人类编写报告的过程。这类方法通过学习原始报告中的潜在变量分布(Li et al., 2023; Tsaniya et al., 2024)，并结合知识图谱等先验知识(Li et al., 2019; Li et al., 2023)，从而生成更加自然和连贯的报告文本。然而，尽管生成式方法在模仿人类写作风格方面取得了一定进展，但它们在生成高价值报告方面仍面临挑战。这是因为，一篇高价值的报告不仅需要信息的准确性和内容的丰富性，更关键的是要具备高度的可解释性和透明的推理过程。报告应当能够清晰地阐述其分析结果、结论和建议的依据，使得读者充分理解报告的核心观点和逻辑基础，从而建立起对报告内容的信任。端到端训练的方法虽然能够捕捉到一定程度的语言规律，但在确保报告的可解释性和推理透明度方面仍有所不足。

随着大型语言模型的出现，大模型凭借其强大的语义理解和生成能力，有望在保持报告内容自然流畅的同时，提供更加详尽的解释和透明的推理路径，生成易于理解和信任的报告。一些工作探究了大模型在自动报告生成上的性能。面对数据量较小的场景下，Ding(2024)直接将有限的原始文本输入到大型模型中，直接生成报告。这种方法简单直接，但真实报告场景下往往需要处理大量文档。这时则倾向于人为规定某些小型模型用于数据处理，如信息提取(Wang et al., 2019)、主题聚类(Wang et al., 2019)等，大模型则根据处理后的数据进行指定内容的推理和生成。这种人为设计的半模板式大模型报告生成方法利用小模型在特定任务上的专业性和高效性，解决了大模型的输入长度问题，但同时也存在一些局限性。首先，这些方法并没有充分发挥大型模型在工具使用和自主规划方面的能力，在生成报告时依然需要依赖人工设定报告模块和对应的小模型类型，这限制了报告内容的全面性和多样性。其次，这种输入提示并直接输出报告的工作流程会由于小模型的错误或冗余的输出和大模型自身生成过程的随机性而不稳定，导致部分报告的质量有所下降。

基于上述存在的两个问题，本文提出了一种大小模型协同的自动报告生成框架AutoRG，旨在通过充分发挥大型语言模型与小型专业模型的互补优势，在确保报告质量的前提下，将人从报告生成的过程中解放（见图1）。AutoRG框架将报告生成任务划分为"大纲规划与评估"、"工具调用与内容修正"和"报告生成与优化"三个阶段。"大纲规划与评估"阶段旨在通过大模型的规划能力，定制化生成和筛选报告所需的子模块，确保每个子模块都能有合适的小模型辅助完成；"工具调用与内容修正"阶段利用大模型的工具理解能力选择性挑选小模型，结合优秀的摘要生成能力对小模型输出进行修正，得到价值密度较高的分析数据；而"报告生成与优化"阶段旨在通过多智能体协作工作流的方式，迭代生成精确、可解释性强且具有深度的报告内容。总结来说，本文的主要贡献可以概括为：


(c) 2024 中国中文信息学会计算语言学专业委员会

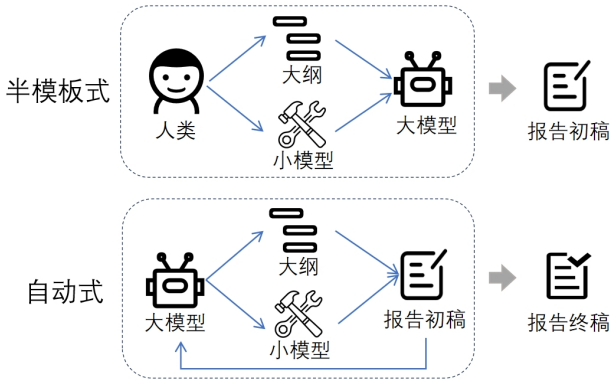



Figure 1: 半模板式与自动式大模型报告生成框架的对比

（1）本文提出了一个大小模型协同的自动报告生成框架AutoRG，有效利用大模型的工具理解和自主规划能力，自动化生成更灵活、全面的报告，无需过多的人工干预。

（2）本文通过引入信息修正和报告迭代机制，降低对小模型输出和大模型随机性的依赖，提高了报告质量的稳定性。

（3）本文以自动专利技术报告生成为例，采用主客观评估结合的方式，从多个维度验证了框架的有效性。

## 2 相关工作

### 2.1 大语言模型

大语言模型（Large Language Model, LLM）是一类在超大规模数据集上进行预训练的大型预训练语言模型，旨在有效地理解和生成人类语言。与小型语言模型相比，大模型因其庞大的参数量和数据训练量，展现出了卓越的涌现能力，特别是在指令跟随和多步推理方面(Zhao et al., 2023)。指令跟随能力允许大模型根据给定的自然语言指令执行任务，即使在缺少具体示例的情况下也能表现出良好的适应性和灵活性。这种能力使得大模型能够在无需额外训练的情况下，理解和执行多样化的任务，如文本生成(Lu et al., 2023)、问题回答(Shi et al., 2023)和代码生成(Chen et al., 2021)等。多步推理能力则使得LLM能够处理涉及多个推理步骤的复杂问题。在适当的提示策略下，大模型能够模拟人类的推理过程，逐步解决如数学问题(Azerbayev et al., 2023)、逻辑谜题(Dua et al., 2022)等任务，显示出对深层次语义和逻辑关系的理解。这些能力使得大模型能够在缺乏直接示例的情况下，根据新任务的指令进行有效的执行，并在处理复杂问题时展现出卓越的性能。

智能体是人工智能领域中的一个核心概念，指能够感知自身所处环境、自我决策并采取行动的人工智能实体(Xi et al., 2023)。大模型的卓越表现在智能体中展现出了巨大的潜力。在大模型智能体框架中，利用大模型强大的零样本指令跟随能力，智能体可以通过根据工具的描述学会如何理解和使用工具，完成更复杂的任务，例如通过搜索工具从外部查询更丰富的信息(Schick et al., 2024)。同时，多步推理能力促进了智能体对复杂任务的规划能力(Xi et al., 2023)。规划使智能体能够将复杂任务分解为子任务(Shen et al., 2024)，并制定合理的行动计划以实现目标。还有一些工作(Hong et al., 2023)合理利用多个智能体之间的协作，使其在各种应用场景中更加有效和可靠。

### 2.2 自动报告生成

#### 2.2.1 抽取式自动报告生成

抽取式报告生成方法通过不同的信息处理小模型获得相关分析结果。这类方法通常需要人工预设好模板，然后使用数据统计(Babour and Khan, 2021)、文本摘要与关键词提取(Wang et al., 2019)等模型，对原始数据进行相关信息的提取，并填入对应的模板中。Babour(2021)提出一个自动生成指定国家科学计量研究分析报告的模型，通过使用数据统计模型对科学出版物数据进行详细的统计分析，填入模板并交由评审员进行评审修改得到分析报告。Noh(2020)提出一种名为Wise Issue Report的自动报告生成系统，利用主题生成模型生成不同主题，通过检索


(c) 2024 中国中文信息学会计算语言学专业委员会

和多文档摘要为每组新闻生成摘要，结合时间事件摘要模型生成不同的摘要内容，组合得到完整报告。

一方面，抽取式报告极度依赖模板，灵活性差；另一方面，抽取式获得的报告只能呈现浅层信息，无法进行深入分析，不适用于专业分析。

#### 2.2.2 生成式自动报告生成

传统的生成式报告生成方法主要通过解码器模型在相关报告数据集上的端到端训练，获得报告生成的能力。Wang(2024)提出了一种基于条件变分自动编码器的方法，使用Bi-GRU编码输入的新闻，通过知识蒸馏和教师-学生网络结构来细化解码器组件的输出，实现金融报告的自动生成。Ren(2021)提出了一种新的混合深度生成神经模型，使用具有注意力机制的指针生成器网络学习输入新闻的大纲，通过改进的变分自动编码器生成宏观金融报告。Li(2023)为了提高放射学报告自动生成的质量，提出知识增强注入框架，通过结合医学概念和相似报告中的临床信息三元组，有效提高了报告的准确性和流畅性。

端到端训练的生成式方法虽然能够使模型掌握一定的报告语言规律并生成连贯的文本，但也往往导致模型过分关注于模仿训练数据中的表面特征，忽视深层语义逻辑。生成的报告可能在形式上看似合理，但在内容深度和逻辑性上却难以达到要求，可解释性不足。此外，高质量的专业报告资源相对稀缺，也限制了其对新领域或新主题报告的泛化能力。

#### 2.2.3 基于大模型的自动报告生成

大模型在自动报告生成领域中正逐渐成为一种趋势。得益于大模型在语义理解与文本生成方面的强大能力，它为生成具有深度和广度的报告内容提供了新的可能性。这类方法结合抽取式和生成式报告生成的优点，采用分工合作的策略，其中大模型主要负责生成连贯的文本内容，而小模型则专注于原始数据的信息抽取，以辅助大模型生成更为详尽和信息丰富的报告。Colverd(2023)提出了一个利用LLM自动生成洪水灾害报告的系统，小模型负责检索关键词相似文档，大模型对文档进行评估并提取与洪水事件相关的信息，通过提示的方式让大模型回答预设的问题，从而生成一个内容全面，结构完整的洪水灾害报告。Reddy(2023)利用大模型自动生成态势报告，使用层次聚类进行新闻分类，利用小型生成模型生成类别标题，使用GPT3基于生成的类别标题进行战略子标题和内容的生成。

这种分工合作的方法虽然能够弥补传统生成式的不足，但仍然存在一些局限性。首先，生成报告的初步阶段通常需要人工设定报告的内容与使用的小模型，这可能会导致报告结构或内容的不全面。例如，在行业趋势预测中，除了行业标签和统计数据外，可能还需考虑龙头企业和专家的发展信息，以增加报告的丰富度和深度。此外，大型语言模型在生成报告时可能会因为回复的不稳定性而产生质量波动。这种不稳定性一方面来自小模型输出的错误与冗余信息，另一方面来自大模型生成文本时的随机性。与人类不同，大模型缺乏自主修正和完善的能力，这也限制了报告内容的优化。

## 3 AutoRG

在本节中，本文将详细介绍AutoRG框架，如图2所示。框架将报告生成的过程分为三个阶段。第一阶段，提供预设的一级题目和可利用的小模型集合，利用大模型生成尽可能丰富且可实现的子模块集合，将其作为报告的二级目录（第3.1节）；第二阶段，利用大模型的工具理解与摘要生成的能力，为每个子模块生成高信息量且更精准的分析数据（第3.2节）；第三阶段，以多智能体协作的方式对报告内容进行迭代优化，获得最终的报告（第3.3节）。

### 3.1 大纲规划与评估

大纲规划阶段的目标是为报告创建一个详尽可实现的目录结构。算法流程如1所示。

**大纲规划**：首先，将预定义的题目模板和可用的小模型集合交给大模型。通过特定的提示，引导大模型为每个题目模板规划生成一系列候选的子模块。这些子模块将作为报告的初始二级目录。在生成子模块时，要求大模型不仅提供模块的标题，还需要预测可能需要的数据类型，以便后续的信息收集。

**可行性评估**：尽管大模型能够生成丰富的子模块集合，但并非所有生成的子模块都能通过现有的小模型集合来实现。因此，需要对这些子模块进行筛选，确保它们与现有的小模型集合相匹配。大模型需要对每个生成的子模块进行反思并做出两个关键决策：确定哪些子模块是可


(c) 2024 中国中文信息学会计算语言学专业委员会

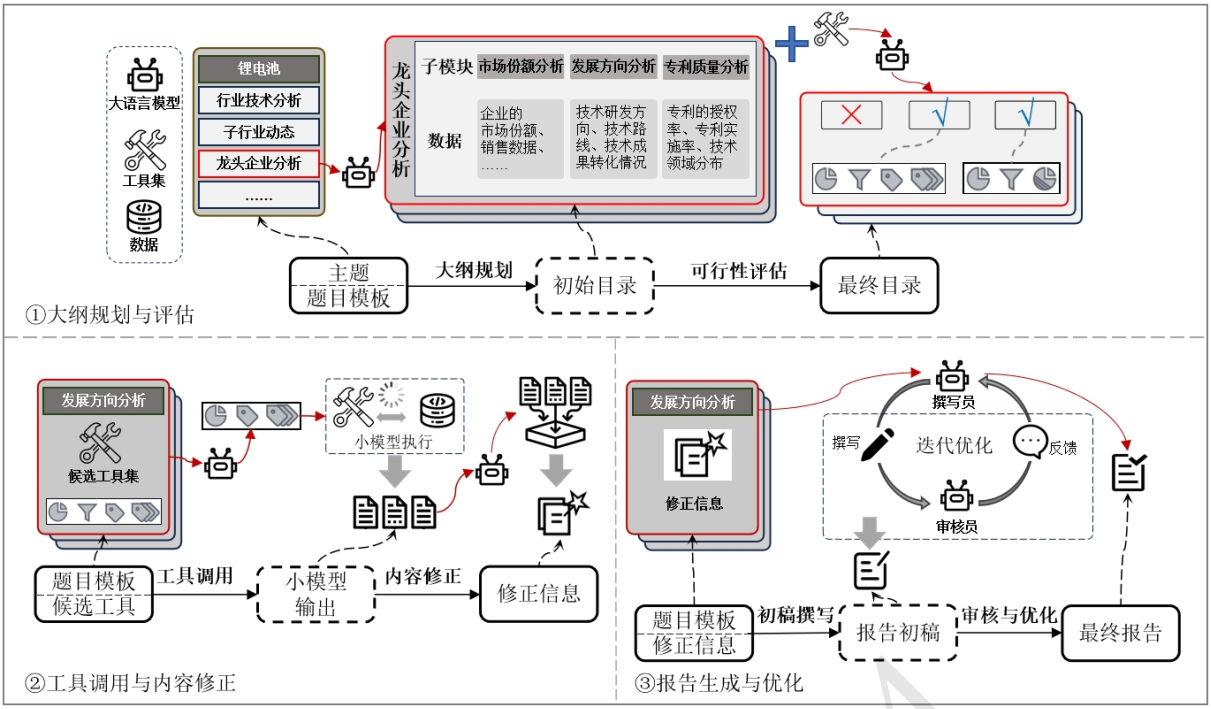



Figure 2: AutoRG报告生成框架

以执行的，以及哪些小模型可能适用于这些子模块。这是小模型首次被引入决策过程，目的是为了简化后续工具调用任务。具体来说，需要向大模型提供了小模型的列表以及完成子模块所需的数据类型。然后，大模型根据提示来判断每个子模块的可执行性，并在必要时从给定的小模型集合中选择合适的小模型子集。如果大模型评估某个子模块无法完成，则该子模块将被排除在最终目录之外。

### 3.2 工具调用与内容修正

工具调用与内容修正阶段旨在大模型完成小模型的自主选择，为每个子模块获取丰富且高质量的分析数据，确保报告的准确性和深度。具体来说，包括以下几个步骤:

**工具调用**：在工具调用步骤中，大模型需要根据当前子模块的任务要求，以及可用的小模型集合，进行精确的调用规划。小模型被视为大模型可利用的工具，大模型需要根据任务的具体需求，理解并决定调用哪些小模型，以获得最佳的输出结果。通过有效的工具选择，使得每个小模型都能在其所擅长的领域内发挥最大的效能，确保信息收集的全面性。

**内容修正**：内容修正的目的是对小模型的输出进行修正优化。大模型在规划步骤之后，并非直接生成报告内容，而是进行信息的过滤和提炼，识别并保留长文本中关键且注意力较大的信息，去除冗余和不相关的部分，降低小模型错误或有干扰的输出信息占比。本文使用直接提示的方式让大模型对小模型的输出进行审查，抽取出最有价值的信息，并将其转化为连贯、流畅的文本段落。这个过程有助于减少信息的冗余，并确保分析数据的高度关联与准确。

### 3.3 报告生成与优化

报告生成与优化部分旨在通过模拟真实人类书写报告时的交互，将修正后的分析数据转化为报告。这一阶段的目标是确保报告不仅表达清晰、逻辑性强，而且信息丰富、分析深入，能够为目标读者提供易于理解且实用的信息。在这一过程中，本文利用大模型分别模拟人类的专家审核员和报告撰写员的角色，通过协作的方式进一步提升报告的质量和准确性。

**初稿撰写**：在生成过程中，审核员将每个子模块的标题和整合后的数据提供给撰写员，要求其构建相应的初稿。在这里，审核员仅指定子模块的标题作为二级标题，而子模块内的报告格式则由撰写员，即由大模型自主决定。这种自主性赋予了大模型灵活性，使其能够根据每个子模块的具体内容来设计最合适的报告结构，确保报告内容的针对性和实用性。


(c) 2024 中国中文信息学会计算语言学专业委员会

**Algorithm 1** 大纲生成算法

**输入:** 题目模板$H = \{h_1, h_2, ..., h_n\}$, 报告主题$domain$, 小模型集合$T = \{t_1, t_2, ..., t_k\}$，大模型$M$
**输出:** 子模块信息矩阵$S = \{s_{ij}\}$，候选小模型信息矩阵$C = \{c_{ij}\}$，其中$s_{ij}$ 是第$i$ 个模板的第$j$ 个子模块，$c_{ij}$ 是$s_{ij}$ 对应的候选小模型列表
**过程:**
**for** $h_i \in H$ **do**
    $S_i = M.generate(h_i, domain)$ ▷ 大纲规划
    **for** $s_{ij}$ in $S_i$ **do**
        **if** $M.judgement(s_{ij})$ **then** ▷ 可行性评估
            $c_{ij} = M.choose(T, s_{ij})$
            保存$s_{ij}$ 与$c_{ij}$
        **else**
            删除$s_{ij}$
        **end if**
    **end for**
**end for**

**Algorithm 2** 工具调用与内容修正算法

**输入:** 第$i$ 个模板的第$j$ 个子模块信息$s_{ij}$，候选小模型列表$c_{ij}$，大模型$M$
**输出:** 优化后的分析数据$data_{new}$
**过程:**
$data_{old} =$ ""
**while** True and $c_{ij} \neq \emptyset$ **do**
    $c_k = M.choose(c_{ij}, s_{ij})$ ▷ 工具调用
    **if** 没有选择$c_k$ **then**
        跳出循环
    **end if**
    $output_c = c_k.run()$
    $c_{ij} = c_{ij} \setminus \{c_k\}$
    $data_{old} = data_{old} + output_c$
**end while**
$data_{new} = correct\_func(data_{old})$ ▷ 内容修正

**审核与优化**：正如一篇优秀的报告往往需要经过审核和反复修改，自动报告生成同样需要经历一个细致的迭代过程，以确保最终产出的报告达到高质量标准。在生成初稿之后，为了减少大模型生成内容的不稳定性，将对报告内容进行审核与优化。具体来说，审核员将对撰写员返回的初稿进行细致的检查，并提出改进意见，这些意见可能涉及格式、内容相关度、分析深度等多个方面。撰写员将根据审核员的反馈，在初稿的基础上进行近一步的优化和完善，从而得到最终的、高质量的报告。

## 4 实验

### 4.1 实验设置

本节以专利技术报告的生成为例，具体阐述相关实现细节。

**数据**：由于资源的限制，本文专注于生成2023年的专利技术分析报告，从特定的在线资源[0]中收集了大量2023年专利数据，包括专利标题、摘要、发明人、申请公司、IPC分类号等信息，并存储在数据库中，以便报告生成使用。为了及时捕捉技术信息，本文以双月作为一个生成周期。在这个周期内，报告覆盖了从宏观到微观、从广泛到具体的不同粒度主题，以确保进

[0]https://analysis.iprdb.com/, https://www.uyanip.com/

CCL 2024


(c) 2024 中国中文信息学会计算语言学专业委员会

**Algorithm 3** 报告迭代算法

**输入:** 第$i$ 个模板的第$j$ 个子模块信息$s_{ij}$，修正后的数据$data_{new}$，审核员$A$，撰写员$M$
**输出:** 第$i$ 个模板的第$j$ 个子模块的报告内容$report_{ij}$
**过程:**
$draft = M.generate\_draft(s_{ij}, data_{new})$ ▷ 初稿撰写
$feedback = A.review\_draft(draft)$ ▷ 审核
**if** 没有新的$feedback$ **then** 不做操作
**else**
    $draft = M.optimize\_draft(draft, feedback)$ ▷ 优化
**end if**
$report_{ij} = draft$

行全面而深入的实验。最终，共计生成23篇报告，138个子模块。

小**模型**：本文设计了报告生成中常见的小模型，包括：标签抽取模型、聚类模型、质量评分模型、统计模型。其中，标签抽取模型负责从专利文档中提取公司、专家和行业的相关技术标签，快速识别专利的技术领域和相关性；聚类模型将相似的专利归类到相应的子类别中，有助于在报告中展示专利的分布情况；质量评分模型通过评估专利的权利要求数量、法律状态等因素，使用岭回归模型对专利进行评分；统计模型对专利数据进行量化分析，包括计算专利的IPC分布、专利类型比例等，为报告提供宏观视角。

**大模型**：大模型在自动报告生成中扮演着重要的角色，负责规划和处理从小模型中获得的数据，以及最终生成连贯的报告文本。考虑到实际应用的可行性和成本效益，本文选择智谱AI作为生成报告的大模型。智谱AI基于GLM构建，通过大量的中文数据训练和对齐，是一个在中文自然语言处理领域表现出色的生成式语言模型。

**对比基线**：AutoRG旨在成为一个通用的报告生成框架，不局限于特定类型的报告，其生成的报告目录由大模型动态决定，这导致在报告结构上与现有相关工作不同。同时，现有的报告生成方法通常专注于特定类型的报告，生成流程和结构目录高度定制化，使得跨方法的直接对比变得复杂。因此，本文从现有工作的工作流，即半模板式大模型报告生成方法的角度进行对比。为确保比较的一致性，本文以大纲生成阶段输出的目录作为报告的标题。在半模板式大模型生成方法中，报告内容的生成使用人工判定的方式，选择最适合当前子模块的一个小模型，并将输出直接交给大模型进行报告生成。

**消融**：同时，为验证方法中不同模块的贡献，本文设置了消融实验。分别通过与不进行工具选择、不进行工具调用规划、不做内容修正以及不进行审核优化下生成的报告进行对比，直观分析这四个部分对最终报告质量的影响。

### 4.2 评估方法

考虑到一篇完整的报告内容篇幅可能过长，导致评估过程变得复杂和耗时，本文选择了以模块为单位进行评估。具体来说，根据二级标题将报告划分为多个独立的部分，每个部分都对应一个特定的报告内容。这样的划分可以使得评估过程更加高效。

**主观评估**：主观评价指标体系的构建参考了以往报告生成质量评估(Babour and Khan, 2021)的工作，包括以下四个维度：事实性、一致性、充分性和整体质量。其中，事实性指标旨在评估报告内容的准确性和真实性；一致性指标关注报告内部逻辑的连贯性以及与既定主题的相关性；充分性指标则评估报告内容是否全面，是否存在信息的遗漏或冗余；整体质量指标综合考量报告是否为读者提供了新的见解和知识。围绕上述评价体系，本文分别用模型评估和人为评估的方式从主观层面对报告生成的质量进行评价。

模型评估为每个评价指标设定独立的投票标准，以确保从不同维度对报告的质量进行细致的考量。本文使用Sparkv3.5[1]和Kimi[2]作为评估模型，通过两两比较的方式让评估模型进行投票。为了避免评估过程中位置造成的影响，本文随机安排了待评估段落的顺序。以半模板式报告作为基准，通过计算胜率来评估不同方法相对于半模板式报告的表现。胜率的计算公式

[1]https://xinghuo.xfyun.cn/
[2]https://kimi.moonshot.cn/


(c) 2024 中国中文信息学会计算语言学专业委员会

| | 事实性 | | | 一致性 | | | 充分性 | | | 整体质量 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Spark | Kimi | avg | Spark | Kimi | avg | Spark | Kimi | avg | Spark | Kimi | avg |
| AutoRG | **61.59** | **61.59** | **61.59** | **50.72** | 60.14 | **55.43** | **64.49** | 64.49 | **64.49** | **60.14** | **63.77** | **61.96** |
| w/o 可行性评估的工具选择 | 58.54 | 56.91 | 57.72 | 39.84 | 58.54 | 49.19 | 57.72 | 63.41 | 60.57 | 56.10 | 59.35 | 57.72 |
| w/o 工具调用规划 | 60.98 | 54.47 | 57.72 | 47.15 | 52.85 | 50.00 | 64.23 | 58.54 | 61.38 | 57.72 | 57.72 | 57.72 |
| w/o 自主信息修正 | 60.98 | 60.16 | 60.57 | 49.59 | 60.16 | 54.88 | 59.35 | **65.04** | 62.20 | 55.28 | 62.60 | 58.94 |
| w/o 内容审核迭代 | 53.03 | 60.61 | 56.82 | 43.94 | **60.61** | 52.27 | 53.79 | 62.12 | 57.95 | 49.24 | 62.12 | 55.68 |

Table 1: 不同报告生成方法下胜率的对比

如1，$k$为报告数量,$n_i$为第$i$篇专利比较的次数，也是子模块的数量，$a_i$为被选择的数量。

考虑到完整报告长度与专业性可能导致评估困难，因此本文采取了以模块为单位的整体评价的方式进行人为评估。每个评审员将随机负责3篇报告的对比，根据评分标准进行综合投票。

$$win = \sum_{i=1}^{k} \frac{a_i}{n_i} \tag{1}$$

**客观评估**：为降低主观性对评估结果的影响，本文进一步设置了不同的客观评估指标。这些指标涵盖了词汇多样性、内容联系度、领域独特性和句法复杂性四个方面，旨在从更丰富的维度对生成报告进行客观对照。词汇多样性的评估采用了类符-形符比（Type-Token Ratio, TTR）作为衡量指标。TTR通过计算文本中不同词汇类型（type）与总词数（token）的比率，来量化文本的词汇丰富度。这一指标能够有效反映报告中词汇的使用广度和多样性。

内容联系度和领域独特性的评估则采用了self-BLEU方法。其中，内容联系度旨在衡量一篇报告内部各个模块之间的联系是否紧密，通过比较报告内部模块与其他模块之间的相似性，来衡量报告整体的联系度。领域独特性旨在评估不同领域之间报告与报告的相似性，如果对于同一个模块，不同领域报告之间的内容过于相似，这样的报告在领域深度上有较大的提升空间。

句法复杂性的评估基于句子嵌套深度的计算。通过测量句子中嵌套句法结构的最大深度，可以评估句子的句法复杂度，进而反映句子结构的复杂性，以及生成文本的句法多样性。

$$TTR = \frac{\#\text{不同类型的词（Type）}}{\#\text{全部词（Token）}} \tag{2}$$

$$self - BLEU = exp(-\sum_{i=1}^{n} w_i \cdot \log \frac{C(i)}{Ni}) \tag{3}$$

## 4.3 实验结果分析

### 4.3.1 主观评估

（1）AutoRG生成的报告在内容上更加充分、全面。如表1所示，模型评估结果揭示了AutoRG与半模板式报告方法相比的胜率，其中高亮显示的部分表示AutoRG的平均胜率超过了半模板式大模型报告生成方法。特别是在报告的充分性这一关键指标上，AutoRG展现出了显著的优越性。这归功于AutoRG利用大模型的自主规划能力，能够自主选择并整合多种小模型，生成内容更为全面和详尽的报告，证实了AutoRG框架在增强报告内容丰富度方面的显著效果。

（2）AutoRG生成的报告在质量上更稳定，效果更好。评估结果显示，AutoRG在事实性、一致性以及整体质量方面均超越了传统的半模板式报告方法，在报告质量上有着一定的优势。此外，通过箱型图（见图3）对生成报告的胜率分布进行可视化分析，可以观察到AutoRG在所有评估指标的胜率分布上均显著高于半模板式报告，并且在一致性、充分性和整体质量中均呈左偏分布，这表明在多数情况下，AutoRG能够产生质量较高的报告。这一统计特性凸显了AutoRG在维持报告质量方面的出色能力。

同时，本文将人为评估结果通过网页形式进行了直观展示[3]。具体来说，图4展示了评审员对投票结果的统计分析。在这项评估中，每位评审员均随机针对评估规则对一篇报告的不同模

[3]http://8.130.143.149:5000/


(c) 2024 中国中文信息学会计算语言学专业委员会

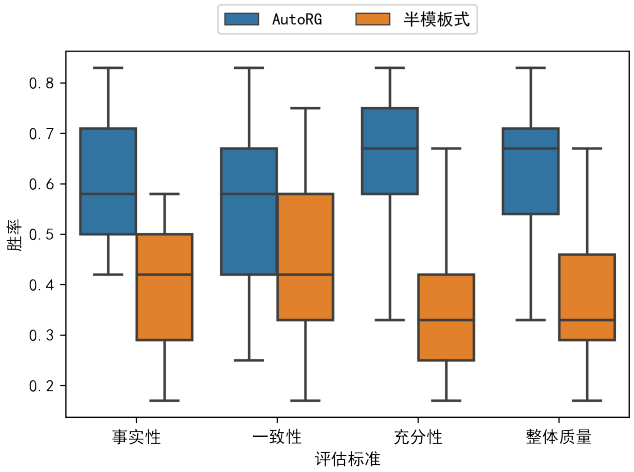



Figure 3: 不同维度下箱型图对比

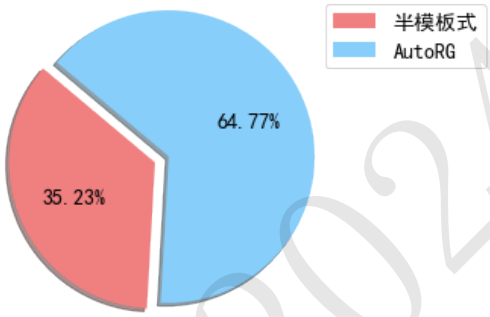



Figure 4: 人为评估投票结果

块分别进行了投票。图中可以清晰地观察到AutoRG在人为评估中获得了压倒性的优势，这一结果进一步证实了AutoRG的有效性和优越性。

### 4.3.2 消融实验

为了更清晰对比不同模块对最终报告生成结果的不同，本文在表2统计了消融实验下各版本与AutoRG平均胜率的差值，并有如下几点结论：

（1）多轮工具选择对提升报告一致性有积极作用。在可行性评估阶段，工具选择和工具调用规划对一致性的影响尤为关键。直接使用所有小型模型作为候选工具集会导致一致性下降6.24%，而将初轮工具选择的结果直接作为候选工具集则导致5.43%的下降。这归结为对庞大工具集的单轮选择虽然一定程度确保了信息的多样性，但也降低了信息的联系度，削弱了子模块与分析内容之间的相关性。而本文采用先粗后精的方式，通过初步筛选和基于详细描述的进一步筛选，提升了报告的一致性和连贯性。

（2）输入数据的质量影响报告的整体质量。自主信息修正模块的作用是对小模型的原始输出做信息的过滤与整合。从表2的结果可以发现，使用初始小模型的输出作为输入数据会导致报告在分析层面的质量下降，这一点在报告整体质量的显著降低中得到了体现。在这一过程中，过多的非相关信息会干扰分析过程，造成输入的冗长和不集中。自主信息修正模块可以进行有效的数据净化，提高报告的分析深度。

（3）迭代对报告内容的优化至关重要。与其他模块相比，当取消内容审核迭代阶段，生成报告的四个指标均有更显著的下降，这揭示了迭代对报告内容优化的重要性。迭代提供了精炼报告内容的机会，允许对初版中可能被忽视的细节和未充分探讨的分析进行深入审视。这种修正和完善的过程，确保了最终报告能够全面且准确地反映分析结果，提高了报告的质量。

第二十三届中国计算语言学大会论文集，第539页-第552页，太原，中国，2024年7月25日至28日。
卷1：主会论文
(c) 2024 中国中文信息学会计算语言学专业委员会

| | 事实性 | 一致性 | 充分性 | 整体质量 |
|---|---|---|---|---|
| AutoRG | - | - | - | - |
| w/o 可行性评估的工具选择 | ↓3.87 | **↓6.24** | ↓3.92 | ↓4.24 |
| w/o 工具调用规划 | ↓3.87 | ↓5.43 | ↓3.11 | ↓4.24 |
| w/o 自主信息修正 | ↓1.02 | ↓0.55 | ↓2.29 | ↓3.02 |
| w/o 内容审核迭代 | **↓4.77** | ↓3.16 | **↓6.54** | **↓6.28** |

Table 2: 消融实验下胜率的变化

| | TTR | 报告内self-BLEU | 报告间self-BLEU | 句法深度 |
|---|---|---|---|---|
| 半模板式 | 0.4805 | 0.0216 | 0.0889 | 4.7793 |
| AutoRG | 0.6066 | 0.0229 | 0.0830 | 4.7889 |

Table 3: 客观指标下不同报告生成方法的对比

#### 4.3.3 客观评估

（1）AutoRG生成的报告在词汇运用上更加多样。如表3所示，AutoRG在TTR指标上明显超越了传统的半模板式方法。这一结果揭示了AutoRG在词汇选择上的丰富性，更有效地传达了信息。这种词汇多样性的实现主要归功于在工具调用、内容修正与迭代阶段的设计。AutoRG不仅能自主选择与任务最匹配的多样化小模型，还能通过信息修正精准提炼关键信息，运用更为丰富和多变的词汇进行表达。这一过程增强了报告的准确性，也提升了其可理解性。而迭代过程的优化则进一步增强了报告的词汇丰富度与表达能力。

（2）AutoRG生成的报告在模块间建立了更紧密的联系。实验数据显示，AutoRG生成的报告在报告内self-BLEU指标上表现出更高的相似度。这得益于规划与工具调用，使得相同的小模型信息在多个模块之间流畅传递。这意味着报告中的各个模块不再是孤立的信息点，而是相互关联、相互支撑的整体。这样的结构使得读者在阅读报告时能够更容易地跟踪信息的脉络，理解各个部分之间的逻辑关系，从而更全面地把握报告的核心内容。

（3）AutoRG生成的报告在不同领域间展现出更高的适应性与差异性。通过对比报告间self-BLEU指标，可以发现，AutoRG在领域多样性方面具有一定的优势。报告之间更高的差异性表明能够根据不同领域的需求，提供更为精准和独特的分析报告。这种差异性依然得益于信息修正模块与报告迭代过程中的细化调整。通过模拟人类专家的审核与优化过程，自动式生成能够对报告进行定制化调整，提高了报告的专业度，也确保了其在不同领域间的差异性和适用性。

（4）AutoRG生成的报告更具专业性。表3中，通过测量句法深度，可以发现AutoRG生成的报告在句法结构的复杂性上表现出细微的优势。这种相对复杂的句子表达说明AutoRG在分析内容上相比半模板式在内容上更为透彻和深入，使得生成的报告在专业性上可能更胜一筹。

## 5 局限与未来方向

在分析AutoRG报告生成优势的同时，本文也关注到其存在的局限，尤其是在事实性与一致性方面。通过评估模型的反馈，发现针对事实性的评估主要集中在真实性方面，无从判断生成报告的准确性与幻觉程度，对于错误信息的处理也依然有较大的改进空间。随着资源的积累和研究的深入，我们也将在未来的工作中继续开展这一问题。此外，AutoRG在报告生成中采用了固定的模块结构，但在实际报告编写过程中，可能需要对报告的目录进行反复调整。本文计划探索大模型在自动调整报告结构方面的潜力，以更好地适应报告编写的动态性和复杂性，为用户提供更加全面和高质量的报告服务。

## 6 总结

本文提出了一个大小模型协同的自动报告生成框架AutoRG。通过有效利用大模型的工具理解与自主规划能力，自动化实现更灵活和全面的报告。同时，通过引入信息修正机制与报告迭代机制，降低对小模型输出和大模型生成随机性的依赖，提高了报告的质量与稳定性。本文在专利场景下以技术报告生成为例，通过模型主观评估和客观评估结合的方式，从多个维度证明，AutoRG生成的报告有着更优的效果。

第二十三届中国计算语言学大会论文集，第539页-第552页，太原，中国，2024年7月25日至28日。
卷1：主会论文
(c) 2024 中国中文信息学会计算语言学专业委员会

## 参考文献

Albert Lu, Hongxin Zhang, Yanzhe Zhang, Xuezhi Wang, and Diyi Yang. 2023. Bounding the capabilities of large language models in open text generation with prompt constraints. *arXiv preprint arXiv:2302.09185*.

Amal Babour and Javed I Khan. 2021. Automatic Generation of a Metric Report: A Case Study of Scientometric Analytics. *IEEE Access*, 10:3923–3934. IEEE.

Christy Y Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. 2019. Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In *Proceedings of the AAAI conference on artificial intelligence*, 33(01):6666–6673.

Dheeru Dua, Shivanshu Gupta, Sameer Singh, and Matt Gardner. 2022. Successive prompting for decomposing complex questions. *arXiv preprint arXiv:2212.04092*.

Grace Colverd, Paul Darm, Leonard Silverberg, and Noah Kasmanoff. 2023. FloodBrain: Flood Disaster Reporting by Web-based Retrieval Augmented Generation with an LLM. *arXiv preprint arXiv:2311.02597*.

Haochen Shi, Weiqi Wang, Tianqing Fang, Baixuan Xu, Wenxuan Ding, Xin Liu, and Yangqiu Song. 2023. Qadynamics: Training dynamics-driven synthetic qa diagnostic for zero-shot commonsense question answering. *arXiv preprint arXiv:2310.11303*.

Hilya Tsaniya, Chastine Fatichah, and Nanik Suciati. 2024. Automatic radiology report generator using transformer with contrast-based image enhancement. *IEEE Access*. IEEE.

Jingwen Wang, Hao Zhang, Cheng Zhang, Wenjing Yang, Liqun Shao, and Jie Wang. 2019. An effective scheme for generating an overview report over a very large corpus of documents. In *Proceedings of the ACM Symposium on Document Engineering 2019*, pages 1–11.

Junpeng Gong, Wen Ren, and Pengzhou Zhang. 2017. An automatic generation method of sports news based on knowledge rules. In *2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)*, pages 499–502. IEEE.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, and others. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Mingjie Li, Rui Liu, Fuyu Wang, Xiaojun Chang, and Xiaodan Liang. 2023. Auxiliary signal-guided knowledge encoder-decoder for medical report generation. *World Wide Web*, 26(1):253–270. Springer.

Qingqiu Li, Jilan Xu, Runtian Yuan, Mohan Chen, Yuejie Zhang, Rui Feng, Xiaobo Zhang, and Shang Gao. 2023. Enhanced Knowledge Injection for Radiology Report Generation. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2053–2058. IEEE.

Revanth Gangi Reddy, Yi R Fung, Qi Zeng, Manling Li, Ziqi Wang, Paul Sullivan, and Heng Ji. 2023. Smartbook: Ai-assisted situation report generation. *arXiv preprint arXiv:2303.14337*.

Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, and others. 2023. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*.

Takeshi Nakaura, Naofumi Yoshida, Naoki Kobayashi, Kaori Shiraishi, Yasunori Nagayama, Hiroyuki Uetani, Masafumi Kidoh, Masamichi Hokamura, Yoshinori Funama, and Toshinori Hirai. 2024. Preliminary assessment of automated radiology report generation with generative pre-trained transformers: comparing results to radiologist-generated reports. *Japanese Journal of Radiology*, 42(2):190–200. Springer.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2024. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, 36.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and others. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.


(c) 2024 中国中文信息学会计算语言学专业委员会

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2024. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*, 36.

Yujuan Ding, Yunshan Ma, Wenqi Fan, Yige Yao, Tat-Seng Chua, and Qing Li. 2024. FashionReGen: LLM-Empowered Fashion Report Generation. *arXiv preprint arXiv:2403.06660*.

Yunpeng Ren, Wenxin Hu, Ziao Wang, Xiaofeng Zhang, Yiyuan Wang, and Xuan Wang. 2021. A hybrid deep generative neural model for financial report generation. *Knowledge-Based Systems*, 227:107093. Elsevier.

Yunseok Noh, Yongmin Shin, Junmo Park, A-Yeong Kim, Su Jeong Choi, Hyun-Je Song, Seong-Bae Park, and Seyoung Park. 2020. WIRE: an automated report generation system using topical and temporal summarization. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2169–2172.

Ziao Wang, Yunpeng Ren, Xiaofeng Zhang, and Yiyuan Wang. 2024. Generating long financial report using conditional variational autoencoders with knowledge distillation. *IEEE Transactions on Artificial Intelligence*. IEEE.

Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen McAleer, Albert Q Jiang, Jia Deng, Stella Biderman, and Sean Welleck. 2023. Llemma: An open language model for mathematics. *arXiv preprint arXiv:2310.10631*.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, and others. 2023. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*.

**附录A.大纲规划与评估的大模型提示**:

| | |
|---|---|
| **大纲规划** | 如果生成“{domain}专利的技术分析报告”，对于其中的”{outline}”模块，可能包括哪些部分?分别需要哪些数据?<br>要求的输出：1. 分点回答；2. 格式是“1. xx分析 |
| **可行性评估** | # 你的任务是生成一个名为”{domain}专利技术分析报告” 的{task_name}模块。<br><br># 完成该模块必须需要以下数据：<br>{task_need_data}<br><br># 有以下候选模型：<br>{model_info}<br><br># 请判断：<br>1.根据候选模型的输出，是否可以获得完成该模块需要的数据？（{task_need_data}）<br>2. 如果不可以，或者需要超过四个模型，那么请直接告诉我“不可以”。（例如，信息统计模型无法获取研发投入信息）<br>3. 如果可以，你应该如何选择和组合这些模型？请注意，最多只能选择四个模型，如果超过四个，参考第二点。（例如，信息统计模型可以获取专利数量信息）<br><br># 返回的格式有两种：<br>- “不可以。因为无法获取xx信息”<br>- “可以。流程是：1. 使用xxx进行xxx\n2. 使用xxx进行xxx” |

CCL 2024

**附录B.工具调用与内容修正的大模型提示:**

| | |
|---|---|
| **工具调用** | ### 任务<br>工具判断。判断以下是否存在符合条件的工具。如果没有，则直接回答‘无法完成‘；如果有，则请从中选择合适的工具，用于完成当前需求。<br><br>### 需求<br>{domain}技术报告-{title}中包含以下模块：{tool_list}，而其中你需要完成的是{tool_key}模块。<br><br>### 工具<br>{tools_info}<br><br>### 要求<br>1. 选择的工具必须能够提供完成{tool_key}需求的全部或部分需要的数据，可以提供地不完整，但是不能完全无关；<br>2. 选择的时候需要考虑与其他模块间的重合度，根据情况可以舍弃重合度高的非关键工具；<br>3. 选择的时候务必带有客观性和严肃性，不要有情感倾向；<br>4. 请严格遵守工具输出的信息，请注意前后回答、逻辑的一致性；<br><br>### 返回格式<br>1. 需求分析。<br>2. 工具分析。包括分析该工具可以提供什么信息? 这个信息和需求有关联吗?<br>3. 综上所述，我的回答是：“无需再选择” 或者综上所述，我的回答是：“xx模型” |
| **内容修正** | # 任务<br>我将给你一段很长的信息文本，请理解背景并耐心阅读文本内容，然后以长文的形式进行信息的过滤与整合。<br><br># 背景<br>{domain}{tool_key}<br><br># 信息:<br>{text}<br><br># 要求<br>1. 保留的内容和{domain}{tool_key}相关；2. 以长文的形式进行总结，保留的内容尽可能丰富、细致；3. 内容尽可能流畅，不要有任何分点的格式；4. 不要过度忽略重要细节，重要细节可能对后续分析有极大帮助；5. 直接回答整合后的长文即可。 |

CCL 2024

**附录C.报告生成与优化的大模型提示:**

| | |
|---|---|
| **初始撰写** | ### 任务<br>我将给你一些对撰写{domain}{tool_key}有帮助的信息，它们可能比较长，请你耐心阅读并撰写{domain}行业技术分析报告——“{tool_key}”模块的内容。<br><br>### 信息<br>{text}<br><br>### 要求<br>1. 第一步先简要梳理关键信息；2. 第二步重点结合{tool_key}，对这些信息分析彼此联系；3. 最后，扩展思维，围绕{tool_key}进行技术层面的深入思考；4. 以“### {tool_key}”开头，直接输出报告即可。 |
| **审核** | 以下是一段{domain}领域技术报告对{subtitle}相关的内容，请你围绕{subtitle}，阅读并审查内容，返回作为专家2 3条最核心的批改意见。<br><br>### 内容<br>”{agent_1}”<br><br>要求返回格式：“1. 意见1；2. 意见2；3. 意见3” |
| **优化** | user: 初始撰写提示<br><br>assistant: {draft}<br><br>user: {advice}<br>当然这只是我的意见，因为我并不了解原信息的全部内容，你可以根据自己的理解对初稿进行润色，使它质量更高、更流畅。<br><br>assistant: |

CCL 2024

# 基于本体信息增强的人类表型概念识别

祁杰蔚, 罗凌*, 杨志豪, 王健, 林鸿飞
大连理工大学，计算机科学与技术学院，大连，116024
jwqi@mail.dlut.edu.cn,{lingluo,yangzh,wangjian,hflin}@dlut.edu.cn

## 摘要

从文本中自动识别人类表型概念对疾病分析具有重大意义。现存本体驱动的表型概念识别方法主要利用本体中概念名和同义词信息，并未充分考虑本体丰富信息。针对此问题，本文提出一种基于本体信息增强的人类表型概念识别方法，利用先进大语言模型进行数据增强，并设计本体向量增强的深度学习模型来提升概念识别性能。在GSC+和ID-68两个数据集上进行实验，结果表明本文提出方法能够利用本体丰富信息有效提升基线模型性能，取得了先进结果。

**关键词：** 生物医学概念识别 ； 人类表型本体 ； 本体信息增强

# Ontology Information-augmented Human Phenotype Concept Recognition

**Jiewei Qi,Ling Luo*, Zhihao Yang, Jian Wang, Hongfei Lin**
School of Computer Science and Technology,Dalian University of Technology,Dalian,116024
jwqi@mail.dlut.edu.cn,{lingluo,yangzh,wangjian,hflin}@dlut.edu.cn

## Abstract

Automatic phenotype concept recognition from text is of great significance for disease analysis. Existing ontology-driven phenotype concept recognition methods mainly utilize concepts and synonyms information within the ontology, without fully considering the rich information of the ontology. To address this issue, this paper proposes an ontology-enhanced method for human phenotype concept recognition, which utilizes advanced large language models for data augmentation and designs a deep learning model enhanced with ontology vectors to improve concept recognition performance. Experimental results on the GSC+ and ID-68 datasets demonstrate that the proposed method effectively leverages the rich ontology information to enhance the performance of baseline models, achieving state-of-the-art results.

**Keywords:** Biomedical Text Mining , Human Phenotype Ontology , Ontology Information Enhancement

## 1 引言

* 通讯作者
©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版
**基金项目**：国家自然科学基金（62302076）；中央高校基本科研业务费（DUT23RC (3)014）；SMP-智谱AI大模型交叉学科基金（ZPCG2024010204)

表型是基因型和环境共同作用的生物体的特定物理外观，是疾病背后生理病理过程的表现，了解表型和疾病之间的关联有助于重大疾病的诊断和治疗。人类表型本体(Human Phenotype Ontology, HPO)全面收集和定义了人类疾病的表型特征，并提供了一套标准化词汇用于描述人类表型异常，在世界上被广泛使用 (Robinson et al., 2008; Köhler et al., 2019; Köhler et al., 2021)。从海量生物医学文本中自动识别出HPO中的表型概念，对于理解人类发病复杂原因，推动生物医学研究的发展，以及提高罕见疾病的临床诊断准确率等方面都具有重要的意义 (Groza et al., 2015)。

人类表型概念识别任务旨在从非结构化的生物医学文本中自动抽取表型概念，如表1所示，该任务包括命名实体识别(Named Entity Recognition, NER)和实体标准化(Named Entity Normalization, NEN)两个子任务，即从文本中识别出人类表型实体并映射到HPO中的标准术语上(HPO ID号)。但目前没有充足的表型概念识别人工语料库资源用来训练基于有监督学习的NER和NEN模型进行流水线识别。因此，大多现存表型概念识别工具主要采用基于词典的方法，例如OBO(Open Biological and Biomedical Ontologies) (Taboada et al., 2014)、NCBO(National Center for Biomedical Ontology) (Jonquet et al., 2009)、Doc2hpo (Liu et al., 2019)等。这些方法通过利用HPO中的概念标准名和提供的同义词构造词典，然后设计基于文本匹配的方法进行HPO概念识别。这类方法简单易实现，并获得了较高的识别准确度，但无法识别词典中不存在的表型概念同义词，存在召回率低的问题。

| | | | | |
|---|---|---|---|---|
| 输入文本 | A syndrome of brachydactyly (absence of some middle or distal phalanges), aplastic or hypoplastic nails, symphalangism... | | | |
| 概念识别 | 14 | 27 | brachydactyly | HP:0001156 |
| | 29 | 71 | absence of some middle or distal phalanges | HP:0009881 |
| | 73 | 103 | aplastic or hypoplastic nails | HP:0001798 |
| | ... | ... | ... | ... |

Table 1: 人类表型概念识别样例

近年来，针对基于词典方法召唤率低的问题，研究者们提出了本体驱动的深度学习方法来识别表型概念，例如NeuralCR (Arbabi et al., 2019)、PhenoBERT (Feng et al., 2022)和PhenoTagger (Luo et al., 2021)等。这类方法将NER和NEN两个子任务转换成一个文本分类任务，通过使用HPO中每个概念的标准名称和同义词构建远程监督数据集，用来训练深度学习分类模型。识别阶段将输入文本转化为$n$元词作为概念候选，并将其输入训练好的分类器进行HPO标识符分类，高于一定阈值的候选被识别为表型概念。这类方法不依赖于人工标注的数据集，同时利用深度学习模型有效提升了概念识别的召回率，在多个表型概念识别测试集上获得了目前先进结果。

上述本体驱动的深度学习方法虽然取得了先进结果，但多数方法主要利用了HPO中的概念名和提供的同义词，其他本体信息并未得到充分利用。而在HPO中，不仅包含了表型概念的同义词信息，还提供了概念定义、概念的上下位关系等丰富信息，这些信息都有助于人类表型概念识别任务。如图1所示，HPO本体结构为树状层次结构，存在上下位关系的术语通常具有十分密切的关联信息。例如“Neoplasm by histology”是其父节点“Neoplasm”更具体的一种表型，这类层次结构信息对于模型识别和区分相似表型概念有重要作用。此外，HPO中多数概念只提供了一个同义词，甚至没有提供同义词。在基于远程监督的深度学习方法中，同义词能够提供更丰富的语义特征，有助于提升分类模型的性能，因此如何对本体中没有同义词信息的概念进行数据增强也值得进一步研究。

最近，大语言模型(Large Language Models, LLMs)在自然语言处理领域的表现引起了广泛的关注，它具有良好的自然语言生成能力、上下文理解能力和知识融合的能力，在处理复杂的自然语言任务方面如命名实体识别、关系抽取、机器翻译等任务上取得优异的性能 (Wang et al., 2023b; Zhou et al., 2023; Wang et al., 2023a)。在一些例如医学、法律、金融等特定领域 (Luo et al., 2024; Cui et al., 2023; Chen et al., 2023)，大模型也可以生成高质量、专业化的内容。目前LLMs在人类表型概念识别任务上处于初步探索阶段，Groza (2024)等人使用GPT-3.5-turbo和GPT-4模型通过构建提示模板对人类表型概念进行抽取，但存在资源耗费大的问题，且大模型生成概念位置和HPO标签ID并不准确，识别结果低于传统方法。


(c) 2024 中国中文信息学会计算语言学专业委员会

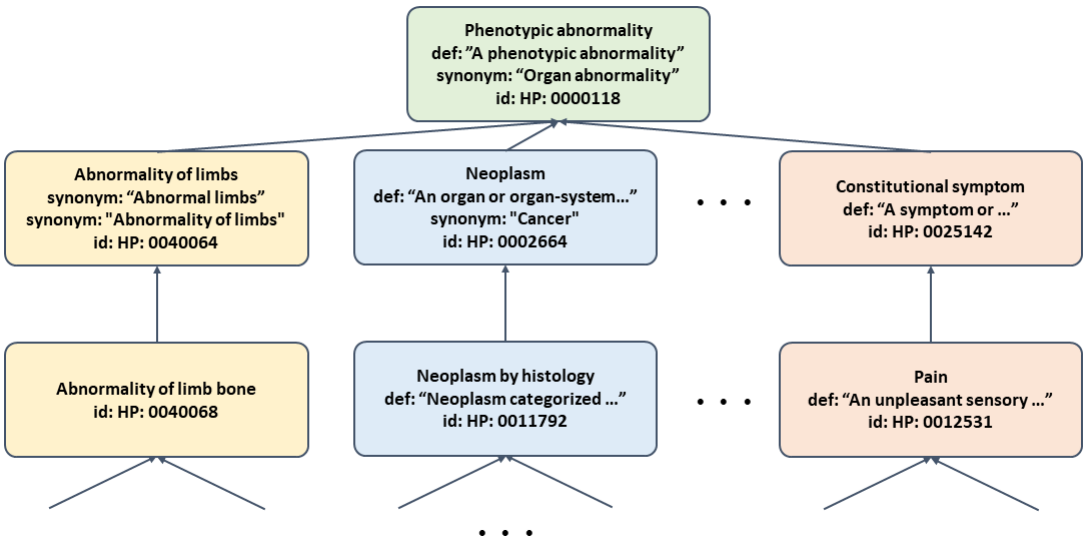



Figure 1: HPO中的信息结构

针对这些问题，本文提出一种基于本体信息增强的人类表型概念识别方法。该方法在先进的本体驱动方法的基础上，借助大语言模型的文本生成能力，利用丰富的HPO本体信息探索了数据增强和模型增强两方面来提升人类表型概念识别的性能。在数据增强方面，利用当前先进的大语言模型GPT-4产生概念术语同义词来增强远程监督训练集；在模型增强方面，通过知识表示方法将HPO的本体结构信息转换为本体向量，用其作为先验知识融入深度学习模型提升模型性能。本文的主要贡献总结如下：

(1) 利用HPO中概念名、概念定义等信息设计大语言模型的提示模板(Prompt)，借助目前先进的大语言模型GPT-4为HPO中缺乏同义词的概念生成其同义词，进行远程监督训练集的数据增强。

(2) 提出了一种融合本体结构向量的深度学习模型对表型概念进行识别。通过使用知识表示学习方法将HPO层次结构信息转换为本体向量，将其与预训练语言模型学习到的文本特征进行融合，提升模型性能。

(3) 在GSC+和ID-68两个表型概念识别数据集上进行实验验证，结果表明本文提出的数据增强和本体向量增强均能提升基础模型性能，两者融合后概念识别性能得到了进一步提升，在两个数据集上取得了最高的平均F1值0.805。这说明了本文方法能够更有效地利用本体丰富信息，验证其有效性和鲁棒性。

## 2 相关工作

最近，人类表型概念识别任务受到了广泛的关注，但由于人工标注语料库稀缺以及表型概念表达多样等问题，人类表型概念识别仍然是生物医学文本处理中的一项具有挑战的任务。现存人类表型概念识别工具主要基于词典匹配方法，包括OBO (Taboada et al., 2014)、NCBO (Jonquet et al., 2009)、Doc2hpo (Liu et al., 2019)、MI(Monarch Initiative) (Shefchek et al., 2020)等。这些基于词典的方法主要是利用HPO构造词典，并通过文本匹配技术对表型概念进行识别。尽管这些方法在精确度上表现出色，但是无法识别出词典中未包含的的概念同义词，存在召回率低的问题。近年来，随着深度学习技术的快速发展，研究者们提供了本体驱动的深度学习方法来进行表型概念识别。由于缺乏人工标记数据集，这些方法大多使用HPO构建远程监督数据集训练深度学习模型进行识别。其中NeuralCR (Arbabi et al., 2019)方法使用稀疏矩阵表示两个概念节点之间的上下位关系，然后融入卷积神经网络(Convolutional Neural Network, CNN)进行概念识别，有效提升了识别召回率，但该方法无法抽取重叠实体概念。PhenoBERT (Feng et al., 2022) 进一步根据本体的层次结构，设计训练了分层CNN结构，能够更准确地识别表型概念。PhenoTagger (Luo et al., 2021)是一种词典和深度学习的混合方法，在词典基础上结合了先进的预训练语言模型BioBERT，同时考虑了重叠概念的抽取，在人


(c) 2024 中国中文信息学会计算语言学专业委员会

类表型概念识别任务上取得了先进结果，但该方法仅利用了HPO中的概念名和同义词信息。本文方法在先进PhenoTagger基础上，重点研究利用丰富本体信息提示模型性能。

为了更充分利用本体结构信息，本文研究了知识表型学习方法来融合本体结构信息。知识表示学习方法旨在将知识库中的实体和关系嵌入到低维稠密的实值向量空间中，如TransE (Bordes et al., 2013)、TransR (Lin et al., 2015)、ConvE (Dettmers et al., 2018)和SemNE (Wang et al., 2021)等。TransE是一种用于知识表示学习的模型，将实体和关系使用向量分布式表示，通过调整向量以近似表示头实体加上关系等于尾实体这一过程。虽然这种方法简单高效，但无法处理复杂关系。为了解决这个问题，TransR被提出，它将实体和关系嵌入到不同的语义空间中，以更好地捕捉复杂相关性。Trans系列模型是浅层模型，简单且参数较少，但在抓取复杂信息方面有所限制。ConvE是一深度学习模型，融合了卷积神经网络和图嵌入技术将实体和关系嵌入到低维向量中，利用卷积操作捕获实体和关系之间的信息。SemNE是一种语义网络编码器，通过学习特征映射函数，并以预训练方式应用于多路语义网络，联合建模语义信息和网络拓扑结构，丰富了网络表示，提高了语义网络的表达能力。

## 3 本体信息增强的表型概念识别方法

本文方法构建了本体信息增强的深度学习标注器。在深度学习标注器中，首先利用HPO构建远程监督训练集，并探索了使用GPT-4产生同义词进行数据增强，然后用其训练本体向量增强的表型概念分类模型，最后对输入文本产生n元候选通过训练好的分类器得到深度学习的识别结果。此外，构建了词典标注器进行结果集成，得到最终的表型概念识别结果。

### 3.1 深度学习标注器

#### 3.1.1 大语言模型增强的远程监督训练集构建

由于表型概念识别人工语料库缺乏，本体驱动方法主要利用HPO中每个概念名、同义词及对应的概念ID，构建远程监督训练集。对于每条概念，将其概念名和其同义词分别产生<**X**:术语文本, **Y**:HPO ID>实例对作为训练集的正例，同时，从生物医学文献中构造一些不属于任何已有概念的负例。即从海量无标注生物医学文本中随机抽取$n$个连续单词，若不在HPO中，则为其分配新的标签“HP:None”作为负例。在构建远程监督训练集的过程中，本文对HPO本体中的概念进行统计分析，结果如图2所示。其中超过40%的概念未提供同义词信息，其余8424条概念包含了同义词，但这之中约50%的概念只包含一条同义词信息。同义词信息的缺乏致使这些概念在训练集中仅有一条样本，从而限制了模型对它们语义信息的学习能力。

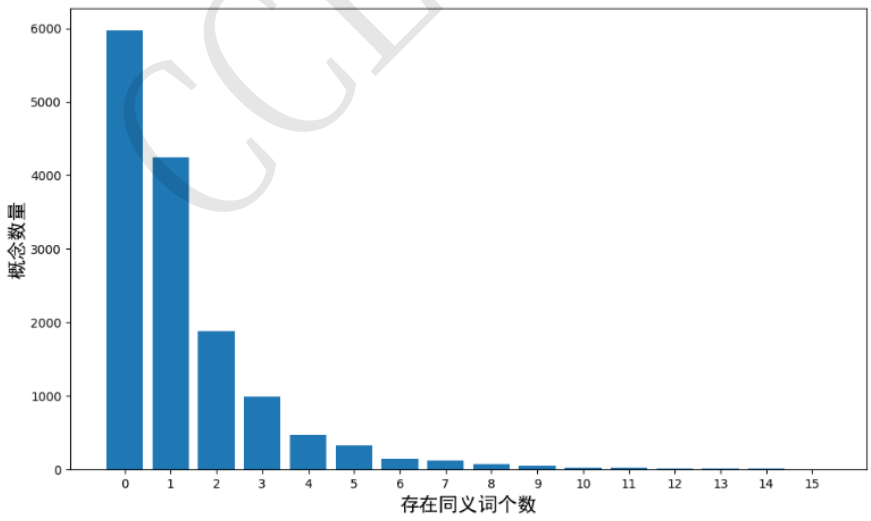



Figure 2: 人类表型概念同义词数量分析

为了缓解同义词信息不足的问题，本文采用了先进的大语言模型GPT-4进行数据增强。先前研究已表明 (Shanahan et al., 2023)，通过设计详细的提示工程(例如，专业领域专家的角色扮演、语义完整的提示模板、提示样例、思维链等)可以让大语言模型显著提升其在该领域


(c) 2024 中国中文信息学会计算语言学专业委员会

的表现。因此，本文在Prompt设定中将大语言模型设定为医学专家的角色，并通过类似思维链 (Wei et al., 2022)的方式，逐步分析概念定义中涉及的症状、病因等要素，并结合人类表型本体的层次结构信息，以及同义词与原始术语之间的语义一致性，通过具体示例来激发大语言模型的理解能力。具体Prompt请参考**附录A.同义词生成Prompt构建**。

### 3.1.2 本体向量增强的表型概念分类模型

基于Transformer (Vaswani et al., 2017)的预训练语言模型，如BERT (Devlin et al., 2018)等，在自然语言处理领域中的各项任务上，相比较于词向量模型、循环神经网络和长短期记忆网络等传统方法有了性能上的提升。本文选择使用生物医学版本的预训练语言模型，如BioBERT (Lee et al., 2020)、Bioformer (Fang et al., 2023)和PubMedBERT (Gu et al., 2021)等作为文本编码器。

本文在生物医学预训练模型基础上，利用知识表示方法获取本体向量，提出一种本体向量增强的表型概念分类模型，模型结构如图3所示。与原始的BERT类似，BioBERT等三种预训练模型也使用WordPiece (Wu et al., 2016)作为嵌入方式，将远程监督训练集中的$\mathbf{X}$作为输入，通过预训练模型编码后得到其隐层向量。具体而言，以输入文本$x_i = (w_1, w_2, \ldots, w_n)$，在每个输入前插入一个特殊标签[**CLS**]得到输入文本序列$x_i^{in}$:

$$x_i^{in} = \{[\mathbf{CLS}], w_1, w_2, ..., w_n\} \tag{1}$$

然后使用预训练模型对$x_i^{in}$进行文本特征提取，采用[**CLS**]对应的向量作为输入文本表示$v_i$。其计算公式如下:

$$v_i = \mathbf{PreTrainModel}(x_i^{in}; \mathbf{W}_0) \tag{2}$$

其中，**PreTrainModel**表示使用预训练语言模型进行编码的过程如BioBERT等,$\mathbf{W}_0$为预训练语言模型的可训练参数。

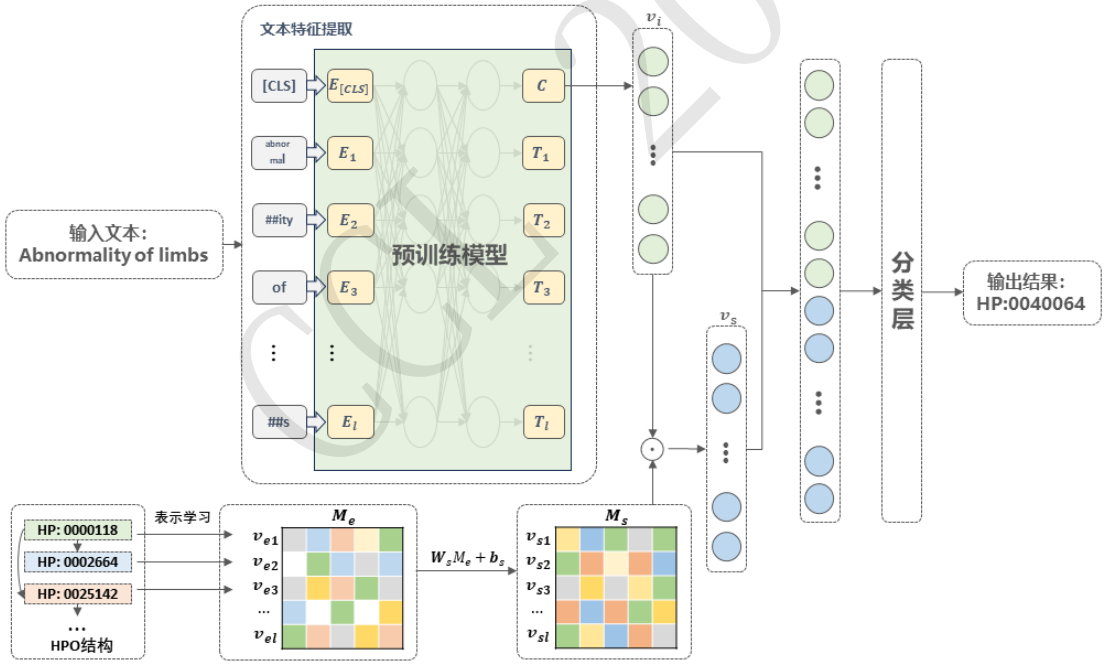



Figure 3: 本体向量增强模型结构图

在本体向量构建部分，本文采用了四种不同的知识表示学习方法对HPO结构信息进行表示。用$T$={$V$,$E$}表示HPO树状结构，其中$V$表示HPO中的概念节点，$E$表示概念节点之间的边。通过知识表示学习方法对HPO树$T$提取结构嵌入矩阵$M_e$。在$M_e$中，每一行表示一个概念实体的嵌入向量。具体过程如下:

$$M_e = \mathbf{RL}(T; \mathbf{W}_1) \tag{3}$$


(c) 2024 中国中文信息学会计算语言学专业委员会

其中，**RL**代表使用知识表示学习方法提取结构矩阵的过程，如TransE等，$W_1$是训练过程中的可训练参数。然后通过一个自定义的结构信息学习层得到结构嵌入矩阵的信息，得到HPO结构信息矩阵$M_s$，其计算公式如下：

$$M_s = \boldsymbol{W}_s M_e + \boldsymbol{b}_s \tag{4}$$

其中，$\boldsymbol{W}_s$和$\boldsymbol{b}_s$为可学习的参数和偏置。之后将结构信息矩阵$M_s$和输入编码向量$v_i$进行点积运算，得到融合结构信息的向量$v_s = M_s \odot v_i$，其中$\odot$表示点积运算。最后将编码向量和融合向量进行拼接，通过一个分类层进行分类得到最后的分类结果$y_i'$，即：

$$y_i' = Softmax\big(Dense([v_i; v_s]; \mathbf{W}_2)\big) \tag{5}$$

其中$\mathbf{W}_2$为分类层的可学习参数。通过以上操作可以得到深度学习方法的预测结果$y_i'$。

最后，本文采用稀疏交叉熵作为损失函数。其损失计算如下：

$$L(y_i, y_i') = -\frac{1}{N}\sum_{i=1}^{N} y_i log(y_i') \tag{6}$$

其中$y_i$是远程监督数据集中$\mathbf{Y}$的真实标签。

#### 3.1.3 概念识别过程

在识别阶段，首先使用NLTK (Bird et al., 2009)将输入文本分句、分词和词性标注，然后生成句子中所有的$n$元词作为候选概念。接下来，对候选概念使用规则方法进行过滤，即根据词性删除以标点符号、介词和连词开始或结束的候选。接下来使用训练好的预训练模型对每个候选对象进行分类。最后，被分类HPO ID标签且预测分数(即预训练模型预测的标签概率分数$P$)高于阈值$t$的候选项被识别为HPO概念，其中$t$是一个超参数，根据本文方法在开发集上的性能进行选择。

### 3.2 词典标注器与结果合并

在PhenoTagger研究中表明融合基于词典方法和深度学习方法可进一步为了提高识别结果的准确率，本文采用同样策略进行词典方法识别结果融合。首先使用HPO中的概念名和同义词构造词典，采用Trie树数据结构 (Fredkin, 1960)来存储HPO词典，进行精确的字符串匹配识别概念。得到词典匹配和深度学习识别的结果后，根据下面四条规则和候选结果得分(词典结果赋予1分，深度学习方法识别结果分值为模型分类概率分值)来获得最终识别结果：1)保留所有不重叠的概念；2)如果重叠的概念具有相同的概念ID，则保留得分最高的概念；3)如果重叠概念在文本中起始和结束位置相同，但被映射到不同的概念ID中，则保留得分最高的概念；4)如果重叠的概念在文本中有不同的开始或结束位置，不同的概念ID，则所有重叠的概念都被保留。

## 4 实验结果与分析

### 4.1 数据集和实验设置

本文主要使用HPO Gold Standardized Corpus plus (GSC+) (Lobo et al., 2017)数据集和ID-68 (Feng et al., 2022)数据集进行实验。随机选择GSC+数据集的10%作为开发集用于超参数选择，其余的数据用作测试集。ID-68数据集包括68份真实的临床记录 (Anazi et al., 2017)，数据集统计结果如表2。其中GSC+数据集规模相对较大，用作主数据集测试方法各模块有效性，将GSC+开发集上设置好的模型直接在ID-68数据集上进行测试。

| | 医学文本数 | 概念数量 | 唯一概念数量 |
|---|---|---|---|
| GSC+ | 228 | 2122 | 1433 |
| ID-68 | 68 | 866 | 800 |

Table 2: 数据集统计

本文采用在开发集上进行随机搜索的方式选择超参数，其中学习率设置为$5 \times 10^{-6}$，batch size大小设置为64，$n$元词的最大长度设置为10，深度学习标注器的分类阈值$t$设置为0.95，并根

据HPO中概念最大长度将$n$元词的最大长度设置为10。使用概念识别任务上广泛采用的评价指标即文档级别的宏平均的Precision，Recall和F1值作为衡量指标 (Luo et al., 2021)。文档级别只需考虑每个文档中的标签集合而忽略文本中的确切位置。

### 4.2 大语言模型数据增强的实验结果

为了验证使用大语言模型增强远程监督训练集的有效性，本节考察了在仅使用GPT-4增强训练集而不改变模型结构的情况下对人类表型概念识别性能的影响。针对此目的，本节中对比三种基座模型BioBERT (Lee et al., 2020)、Bioformer (Fang et al., 2023)和PubMedBERT (Gu et al., 2021)的性能。本文使用GPT-4对不包含同义词的术语概念产生3个同义词，经过实验对比证明在使用2个同义词时性能最优，详细对比结果请参考**附录B.同义词数量在GSC+数据集上对模型的性能影响**。下述实验均使用2个同义词进行模型性能对比。使用GPT-4对HPO同义词增强后在GSC+测试集上的性能对比结果如表3所示，其中Dict表示使用词典匹配方法，DL表示深度学习(括号中为不同基座模型下的结果)，Hybrid为词典匹配方法和深度学习方法结合的结果(括号中为不同基座模型下的结果)。

| 方法 | 未用GPT-4增强 | | | 使用GPT-4增强 | | |
|---|---|---|---|---|---|---|
| | Precision | Recall | F1 | Precision | Recall | F1 |
| Dict | 0.756 | 0.565 | 0.647 | 0.765 | 0.587 | 0.664 |
| DL(BioBERT) | 0.754 | 0.700 | 0.726 | 0.762 | 0.736 | 0.749 |
| DL(Bioformer) | 0.767 | 0.621 | 0.686 | 0.743 | 0.700 | 0.721 |
| DL(PubMedBERT) | 0.755 | 0.706 | 0.730 | 0.749 | 0.720 | 0.734 |
| Hybrid(BioBERT) | 0.766 | 0.712 | 0.738 | **0.768** | **0.751** | **0.760** |
| Hybrid(Bioformer) | **0.773** | 0.674 | 0.720 | 0.760 | 0.719 | 0.739 |
| Hybrid(PubMedBERT) | 0.762 | **0.725** | **0.743** | 0.766 | 0.736 | 0.751 |

Table 3: GPT数据增强在GSC+测试集上的结果。其中Dict表示使用词典匹配方法，DL表示深度学习方法，Hybrid表示词典和深度学习混合方法，加粗表示最高值。

实验结果表明：(1) 使用GPT-4生成同义词后，对词典匹配方法的准确率和召回率分别提升了0.9%和2.2%，表明GPT-4生成的同义词中包含了原有词典中不存在的正确概念；(2) 使用GPT-4扩充远程监督训练集也能提高深度学习方法的性能，其中在Bioformer模型上提升最为明显，达到了3.5%，在BioBERT模型上的提升也达到了2.3%；(3)混合词典和深度学习方法后，GPT-4增强的BioBERT模型获得了最高F1值。这说明使用GPT-4生成同义词后，能够改善HPO中缺少同义词概念的识别性能，并且对于原始识别性能较弱的基座模型提升效果更加明显。

### 4.3 本体向量增强的实验结果

为了验证使用本体结构信息增强概念识别的有效性，本节考察了在不使用大语言模型增强训练集的情况下，仅使用本体结构信息增强模型的识别性能。主要探究不同本体向量维度和不同本体表示方法对模型性能的影响。

#### 4.3.1 本体向量维度对模型性能的影响

本实验主要探索不同嵌入矩阵中向量维度对人类表型概念识别的影响，使用SemNE方法作为结构信息嵌入向量生成方法，BioBERT作为基座模型，在GSC+测试集上进行对比，实验结果如图4所示。


(c) 2024 中国中文信息学会计算语言学专业委员会

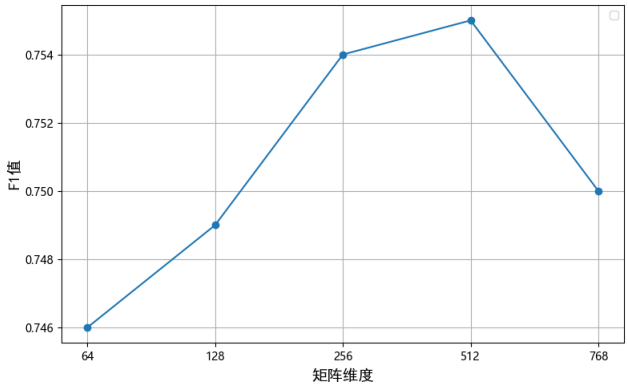



Figure 4: 不同维度的本体向量在GSC+数据集上对模型的性能影响

从实验结果可以看出，随着维度从64增加至512，模型性能持续提升，但在768时性能出现下降。这可能是由于随着向量维度的增加，嵌入向量中包含的结构信息变得更加丰富，但进一步增加维度可能导致一些过拟合情况，从而影响了性能。因此，后续实验选择512作为本体向量的维度。

### 4.3.2 本体向量表示方法对模型性能的影响

本实验主要探索不同知识表示方法对人类表型概念识别的影响，分别使用TransE、TransR、ConvE和SemNE四种方法作为结构信息嵌入向量的提取方法，维度设置为512，同时，测试了BioBERT、Bioformer和PubMedBERT不同的基座模型，以评估不同提取方法的通用性。在GSC+数据集上的实验结果如图5所示，其中“Original”表示不使用嵌入向量，“Random”表示使用随机函数随机生成嵌入向量。

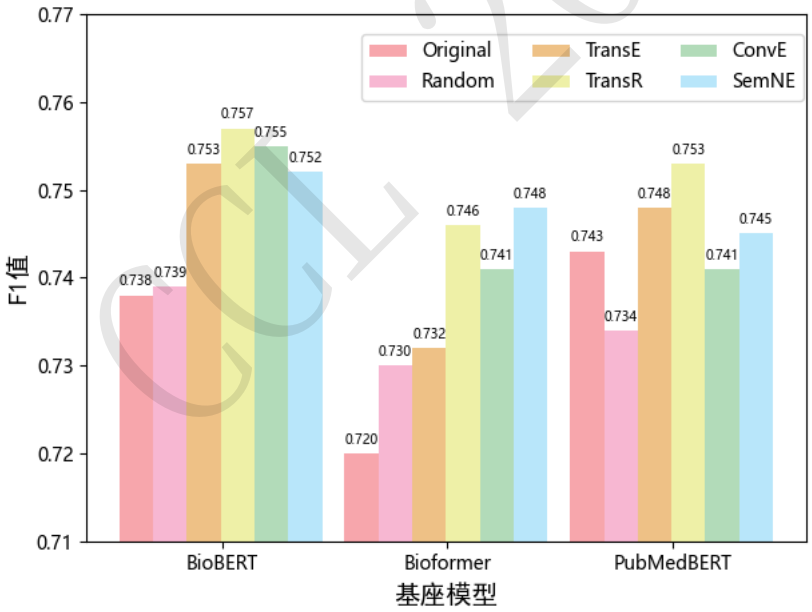



Figure 5: 不同知识表示方法在GSC+数据集上对模型的性能影响

从图中结果可以观察到：(1) 三种基座模型在添加不同方法构建的结构信息嵌入向量后，均提升了人类表型概念识别的性能。特别是在使用BioBERT作为基座模型时，并使用TransR作为结构信息向量提取方法时表现最佳；(2) 在不同模型上，使用TransR作为嵌入向量生成方法获得了最高的平均F1值0.752，SemNE的平均性能也达到了0.748。这表明模型可以从TransR和SemNE提取到的结构信息嵌入向量中学到最丰富的信息；(3) 为了验证使用不同方法提取结构信息向量的有效性，本文使用随机函数随机生成嵌入向量，并将其添加到三种


(c) 2024 中国中文信息学会计算语言学专业委员会

基座模型中。结果显示，相比于其他嵌入向量生成方法，随机生成的嵌入向量在三种基座模型上性能较差，进一步验证了使用不同方法进行结构信息向量提取可以提高表型概念的识别性能。

#### 4.3.3 数据增强方法对模型性能的影响

本实验主要探索不同数据增强方法对模型性能的影响，本文选择基于BioBERT的PhenoTagger作为基线模型（Baseline），增加了三种常用的数据增强方法进行比较：1）SynonymAug：根据WordNet中的距离计算相似度从而选取同义词进行数据增强；2）RandomWordAug：对概念中的词汇随机选取一个其他单词进行替换进行数据增强；3）SpellingAug：引入拼写错误的进行数据增强。Our为本文中使用GPT进行增强的模型。在GSC+数据集上的实验结果如表4所示。

| 方法 | GSC+数据集 | | |
|---|---|---|---|
| | Precision | Recall | F1 |
| Baseline | 0.766 | 0.712 | 0.738 |
| SynonymAug | **0.774** | 0.723 | 0.745 |
| RandomWordAug | 0.295 | 0.661 | 0.407 |
| SpellingAug | 0.754 | 0.725 | 0.739 |
| Ours | 0.768 | **0.751** | **0.760** |

Table 4: 数据集统计

从表4中可以观察到：（1）在使用WordNet同义词（SynonymAug）和拼写错误（SpellingAug）的数据增强时，相对于Baseline分别提升了0.7%和0.1%，但是在使用随机替换（RandomWordAug）的增强方法时，模型性能下降了约30%，其原因可能是随即替换引入了过多的噪声，导致模型无法进行识别；（2）使用GPT-4进行数据增强对比其余几种方法在性能上的提升更加明显，其原因可能是在使用GPT-4产生同义词的时候即考虑了术语概念也同时考虑到了术语的上下位关系，因此可以获得在语义上更加相似的增强数据。

### 4.4 与其他现存方法性能对比结果

前序实验分别验证了利用GPT-4进行数据增强和本体向量增强方法的有效性，本节进一步将两者进行合并，具体使用BioBERT作为基座模型，本体向量表示方法使用TransR，同时加入GPT-4增强的数据进行训练。在GSC+和ID-68两个数据集上将本文方法(Ours)与现有先进方法进行了比较，验证其有效性，结果如表5所示，使用领域内常用的NEN指标进行性能评估 (Luo et al., 2021)，平均F1值由GSC+数据集和ID-68数据集的F1值求平均得到。本文方法在NER和NEN两个子任务上更详细的实验结果请参考**附录C.本文方法详细结果**。

- OBO: 专门用于注释具有HPO表型异常的生物医学文本工具，使用基于词汇和上下文的匹配方法。
- MI(Monarch Initiative) (Shefchek et al., 2020): Monarch Initiative是一个集成数据和分析平台，将不同物种的表型与基因型联系起来，允许用户输入自由文本，并使用Monarch知识图中的术语对文本进行自动表型注释。
- Doc2hpo (Liu et al., 2019): 一个从临床文本中提取人类表型本体的方法，使用了Doc2hpo的集成引擎，该引擎结合了多个基于字符串的方法。
- NeuralCR (Arbabi et al., 2019): 一种基于CNN的人类表型概念识别模型。
- Txt2hpo(github.com/GeneDx/txt2hpo/tree/develop/txt2hpo): 一个用于从文本中提取HPO编码的表型的Python库。
- PhenoBERT (Feng et al., 2022): 一个以BERT为核心，在BERT前引入了一个两层的CNN模块的深度学习方法。


(c) 2024 中国中文信息学会计算语言学专业委员会

- GPT-3.5/GPT-4: Groza (2024)使用不同的Prompt在GPT-3.5和GPT-4上完成人类表型概念识别任务，本文选取两种模型最好的结果作为对比。
- PhenoTagger (Luo et al., 2021): 一种将词典匹配和深度学习方法结合的混合方法。

| 方法 | GSC+数据集 | | | ID-68数据集 | | | 平均F1 |
|---|---|---|---|---|---|---|---|
| | Precision | Recall | F1 | Precision | Recall | F1 | |
| OBO | 0.809 | 0.565 | 0.665 | 0.852 | 0.622 | 0.719 | 0.692 |
| MI | 0.757 | 0.605 | 0.673 | 0.879 | 0.438 | 0.584 | 0.629 |
| Doc2hpo* | 0.768 | 0.618 | 0.685 | 0.859 | 0.580 | 0.692 | 0.689 |
| NeuralCR* | 0.741 | 0.602 | 0.664 | 0.796 | **0.794** | 0.795 | 0.730 |
| Txt2hpo | **0.825** | 0.562 | 0.669 | 0.909 | 0.683 | 0.780 | 0.725 |
| PhenoBERT | 0.813 | 0.640 | 0.716 | **0.946** | 0.776 | **0.852** | 0.784 |
| GPT-3.5* | 0.410 | 0.410 | 0.410 | - | - | - | - |
| GPT-4* | 0.750 | 0.470 | 0.580 | - | - | - | - |
| PhenoTagger | 0.766 | 0.712 | 0.738 | 0.916 | 0.766 | 0.834 | 0.786 |
| Ours(GPT) | 0.768 | **0.751** | 0.760 | 0.912 | 0.780 | 0.841 | 0.801 |
| Ours(EMB) | 0.786 | 0.730 | 0.757 | 0.928 | 0.770 | 0.842 | 0.799 |
| Ours(GPT+EMB) | 0.783 | 0.744 | **0.763** | 0.924 | 0.781 | 0.846 | **0.805** |

Table 5: 与现存方法对比结果。其中GPT表示使用GPT-4扩充数据，EMB表示使用本体向量增强。*表示结果引用于原论文，加粗表示最高值。

从表5结果中可以观察到：(1) 在GSC+和ID-68两个数据集上，相比原始的PhenoTagger，本文提出的本体向量增强和GPT数据增强两种改进方法都取得了性能上的提升，将两种方法合并后获得了进一步的提升，在两个数据集上取得了最高平均F1值，为0.805，相比目前先进的基于深度学习的本体驱动方法PhenoTagger和PhenoBERT在平均F1值上分别提升了1.9%和2.1%，表明了本文方法的有效性和鲁棒性；(2) 在GSC+数据集上，与基于词典的方法相比，本文方法在准确率方面略有降低，但在召回率方面平均提升了近15%。而在ID-68数据集上，相较于基于词典的方法，本文方法在精确度和召回率方面平均都有超过8%的提升。模型在ID-68数据集上的效果优于在GSC+数据集上的效果，其原因可能是ID-68数据集仅包含68篇电子病历，其中概念数量为866，唯一概念数量为800；而GSC+数据集包含228篇生物医学文献摘要，其中概念数量为2122，唯一概念数量为1433。相比GSC+数据集，ID-68数据集规模较小而且表型概念类型较少，识别难度相对较低，因此模型在ID-68数据集上的效果明显优于在GSC+数据集上的效果；(3)在ID-68数据集上，本文方法略低于PhenoBERT方法，其原因可能是不同于数据集GSC+来源于生物医学文献，ID-68数据集由68篇电子病历组成。PhenoBERT在产生概念候选时，使用了Stanza工具中的"ner-i2b2"进行了电子病历实体识别，而本文方法只使用了n-gram产生候选。在电子病历上，PhenoBERT利用额外工具产生了更准确的实体候选，因此在电子病历文本ID-68上，本文方法略差于PhenoBERT（F值相差0.006），但根据实验结果在两个数据集上本文方法获得了最高的平均F1值，说明了本文方法的鲁棒性和泛化能力；(4) 直接使用大语言模型GPT3.5/4进行人类表型识别，结果低于词典方法，这可能受限于目前大语言模型对HPO中概念ID知识学习，概念识别这类任务对先进大语言模型仍具有巨大挑战，而本文使用GPT进行数据增强能够进一步提升模型性能。综上，本文提出的方法能够有效提升原始模型的识别性能，在两个数据集均获得了先进结果。

### 4.5 识别结果样例分析

为了深入验证本文方法的有效性，本文对数据增强最优模型和本体向量增强最优模型结果上进行了人工分析，选取了两个具体案例进行展示。

表6显示了在使用GPT增强远程监督训练集前后对同一条生物医学文本的识别结果。可以观察到，在使用GPT增强后，成功识别出了原有模型遗漏的概念，并且概念的分类和跨度均被正确识别。表7则展示了在引入结构信息前后对同一条生物医学文本的识别结果。在引入结构信息前，模型识别错误的概念，并且分类也出现错误。在当前使用的HPO版本


(c) 2024 中国中文信息学会计算语言学专业委员会

中，“HP:00009371”是“HP:00001156”的孙子节点，在引入结构信息后，成功识别出概念的位置和类别。这两个实例表明了本文方法的可行性。

| 输入文本：“... have orbital haemangiomatous cysts, ...” | | | | |
|---|---|---|---|---|
| 正确答案 | 78 | 107 | orbital haemangiomatous cysts | HP:0001144 |
| 未使用GPT增强 | \ | \ | \ | \ |
| GPT增强 | 78 | 107 | orbital haemangiomatous cysts | HP:0001144 |

Table 6: 使用GPT增强后性能提升样例

| 输入文本：“Brachydactyly type A-1 (BDA1) was, in 1903...” | | | | |
|---|---|---|---|---|
| 正确答案 | 0 | 22 | Brachydactyly type A-1 | HP:0009371 |
| 未引入结构信息 | 0 | 13 | Brachydactyly | HP:0001156 |
| 引入结构信息 | 0 | 22 | Brachydactyly type A-1 | HP:0009371 |

Table 7: 引入结构信息后性能提升样例

## 5 结论与展望

针对目前人类表型概念识别方法未充分使用本体信息的问题，本文提出从数据增强和模型增强两方面利用本体丰富信息提升模型识别性能。在数据增强方面，通过本体信息设计提示模板，利用先进大语言模型GPT-4产生同义词进行远程监督数据增强；在模型增强方面，提出本体向量增强的深度学习模型，通过设计知识表示向量融合本体结构信息提示模型性能。将两种方法融合后，在两个测试集上均取得了性能的进一步提升，验证了方法的有效性。

在未来工作中，我们将深入探索更先进的知识表示方法来充分利用除结构层次信息外的其他丰富信息，例如术语定义、术语评论等。同时，本文方法基于本体驱动的概念识别方法，不依赖于人工标注数据集，未来也将在更多生物医学本体上进行测试。此外，大语言模型展示出了巨大潜力，探索基于大语言模型的概念识别生成式框架也是我们未来工作的方向。

## 参考文献

Shams Anazi, Sateesh Maddirevula, Vincenzo Salpietro, Yasmine T Asi, Saud Alsahli, Amal Alhashem, Hanan E Shamseldin, Fatema AlZahrani, Nisha Patel, Niema Ibrahim, et al. 2017. Expanding the genetic heterogeneity of intellectual disability. *Human genetics*, 136:1419–1429.

Aryan Arbabi, David R Adams, Sanja Fidler, Michael Brudno, et al. 2019. Identifying clinical terms in medical text using ontology-guided machine learning. *JMIR medical informatics*, 7(2):e12596.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. ” O’Reilly Media, Inc.”.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26.

Wei Chen, Qiushi Wang, Zefei Long, Xianyin Zhang, Zhongtian Lu, Bingxuan Li, Siyuan Wang, Jiarong Xu, Xiang Bai, Xuanjing Huang, and Wei Zhongyu. 2023. Disc-finllm: A chinese financial large language model based on multiple experts fine-tuning. *arXiv preprint arXiv:2310.15205*.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *arXiv preprint arXiv:2306.16092*.

Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. 2018. Convolutional 2d knowledge graph embeddings. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.


(c) 2024 中国中文信息学会计算语言学专业委员会

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Li Fang, Qingyu Chen, Chih-Hsuan Wei, Zhiyong Lu, and Kai Wang. 2023. Bioformer: an efficient transformer language model for biomedical text mining. *ArXiv*.

Yuhao Feng, Lei Qi, and Weidong Tian. 2022. Phenobert: a combined deep learning method for automated recognition of human phenotype ontology. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 20(2):1269–1277.

Edward Fredkin. 1960. Trie memory. *Communications of the ACM*, 3(9):490–499.

Tudor Groza, Sebastian Köhler, Sandra Doelken, Nigel Collier, Anika Oellrich, Damian Smedley, Francisco M Couto, Gareth Baynam, Andreas Zankl, and Peter N Robinson. 2015. Automatic concept recognition using the human phenotype ontology reference and test suite corpora. *Database*, 2015:bav005.

Tudor Groza, Harry Caufield, Dylan Gration, Gareth Baynam, Melissa A Haendel, Peter N Robinson, Christopher J Mungall, and Justin T Reese. 2024. An evaluation of gpt models for phenotype concept recognition. *BMC Medical Informatics and Decision Making*, 24(1):30.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Clement Jonquet, Nigam Haresh Shah, and Mark A. Musen. 2009. The open biomedical annotator. *Summit on Translational Bioinformatics*, 2009:56 – 60.

Sebastian Köhler, Leigh Carmody, Nicole Vasilevsky, Julius O B Jacobsen, Daniel Danis, Jean-Philippe Gourdine, Michael Gargano, Nomi L Harris, Nicolas Matentzoglu, Julie A McMurry, et al. 2019. Expansion of the human phenotype ontology (hpo) knowledge base and resources. *Nucleic acids research*, 47(D1):D1018–D1027.

Sebastian Köhler, Michael Gargano, Nicolas Matentzoglu, Leigh C Carmody, David Lewis-Smith, Nicole A Vasilevsky, Daniel Danis, Ganna Balagura, Gareth Baynam, Amy M Brower, et al. 2021. The human phenotype ontology in 2021. *Nucleic acids research*, 49(D1):D1207–D1217.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29.

Cong Liu, Fabricio Sampaio Peres Kury, Ziran Li, Casey N. Ta, Kai Wang, and Chunhua Weng. 2019. Doc2hpo: a web application for efficient and accurate hpo concept curation. *Nucleic Acids Research*, 47:W566 – W570.

Manuel Lobo, Andre Lamurias, and Francisco M Couto. 2017. Identifying human phenotype terms by combining machine learning and validation rules. *BioMed Research International*, 2017.

Ling Luo, Shankai Yan, Po-Ting Lai, Daniel Veltri, Andrew Oler, Sandhya Xirasagar, Rajarshi Ghosh, Morgan Similuk, Peter N Robinson, and Zhiyong Lu. 2021. Phenotagger: a hybrid method for phenotype concept recognition using human phenotype ontology. *Bioinformatics*, 37(13):1884–1890.

Ling Luo, Jinzhong Ning, Yingwen Zhao, Zhijun Wang, Zeyuan Ding, Peng Chen, Weiru Fu, Qinyu Han, Guangtao Xu, Yunzhi Qiu, Dinghao Pan, Jiru Li, et al. 2024. Taiyi: a bilingual fine-tuned large language model for diverse biomedical tasks. *Journal of the American Medical Informatics Association: JAMIA*, pages ocae037–ocae037.

Peter N Robinson, Sebastian Köhler, Sebastian Bauer, Dominik Seelow, Denise Horn, and Stefan Mundlos. 2008. The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. *The American Journal of Human Genetics*, 83(5):610–615.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*, 623(7987):493–498.

CCL 2024

Kent A Shefchek, Nomi L Harris, Michael Gargano, Nicolas Matentzoglu, Deepak Unni, Matthew Brush, Daniel Keith, Tom Conlin, Nicole Vasilevsky, Xingmin Aaron Zhang, et al. 2020. The monarch initiative in 2019: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic acids research*, 48(D1):D704–D715.

Maria Taboada, Hadriana Rodríguez, Diego Martínez, María Pardo, and María Jesús Sobrido. 2014. Automated semantic annotation of rare disease cases: a case study. *Database*, 2014:bau045.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Zhizheng Wang, Yuanyuan Sun, Xuyang Hu, Jiafeng Zhao, Zhihao Yang, and Hongfei Lin. 2021. A semantic network encoder for associated fact prediction. *IEEE Transactions on Knowledge and Data Engineering*, 34(11):5114–5125.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023a. Document-level machine translation with large language models. *arXiv preprint arXiv:2304.02210*.

Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, et al. 2023b. Instructuie: multi-task instruction tuning for unified information extraction. *arXiv preprint arXiv:2304.08085*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2023. Universalner: Targeted distillation from large language models for open named entity recognition. *arXiv preprint arXiv:2308.03279*.

CCL 2024

## 6 附录

### A.同义词生成Prompt构建

本文通过使用先进的大语言模型进行数据增强来缓解同义词信息不足的问题。通过设计详细的Prompt激发大语言模型的理解能力来生成同义词。本文在实验过程中发现HPO术语中有一部分术语具备明确的定义，而另一部分术语并未包含明确的定义。因此，在构建Prompt的过程中，本文分别针对这两种情况进行了处理。对于存在术语定义的概念，本文首先设定大语言模型扮演医学专家的角色，然后对定义中的症状、病因等因素进行分析，结合HPO的层次结构信息以及同义词与原始术语的语义一致性，通过示例构建同义词。而对于不包含明确定义的概念，则无需进行定义的分析步骤。具体Prompt如表8。

| | Prompt |
|---|---|
| 有定义概念 | As a medical expert, you will be given the name of a human phenotype term along with its definition and subclassing relationships. Your task is to generate synonyms for the phenotype by following these steps:<br>1. Analyze the symptoms, causes and other significant factors in the definition of the phenotype;<br>2. Differentiate the term from its parent term by utilizing subclass relationships;<br>3. Ensure that the synonyms maintain the contextual meaning of the original name;<br>4. Provide the top three synonyms, ranked from high to low based on their similarity to the term. No need to include detailed analysis steps in the response.<br><br>##<br>Example:<br>Input:<br>Name: "Multicystic kidney dysplasia"<br>Definition: "Multicystic dysplasia of the kidney is characterized by multiple cysts of varying size in the kidney and the absence of a normal pelvicaliceal system. The condition is associated with ureteral or ureteropelvic atresia, and the affected kidney is nonfunctional."<br>Subclass relationship: "Multicystic kidney dysplasia is a subclass of Renal cyst."<br><br>Response:<br>1. "Multicystic renal dysplasia"<br>2. "Multicystic dysplastic kidney"<br>3. "Multicystic kidneys"<br>## |
| 无定义概念 | As a medical expert, you will be given the name of a human phenotype term and its subclassing relationships. Your task is to generate synonyms for the phenotype by following these steps:<br>1. Differentiate the term from its parent term by utilizing subclass relationships;<br>2. Ensure that the synonyms maintain the contextual meaning of the original name;<br>3. Provide the top three synonyms, ranked from high to low based on their similarity to the term. No need to include detailed analysis steps in the response.<br><br>##<br>Example:<br>Input:<br>Name: "Multicystic kidney dysplasia"<br>Subclass relationship: "Multicystic kidney dysplasia is a subclass of Renal cyst."<br><br>Response:<br>1. "Multicystic renal dysplasia"<br>2. "Multicystic dysplastic kidney"<br>3. "Multicystic kidneys"<br>## |

Table 8: 同义词生成Prompt构建

CCL 2024

第二十三届中国计算语言学大会论文集，第553页-第567页，太原，中国，2024年7月25日至28日。
Proceedings of the 23rd China National Conference on Computational Linguistics, pages 553-567, Taiyuan, China, July 25-28, 2024.
(c) 2024 中国中文信息学会计算语言学专业委员会

**B.同义词数量在GSC+数据集上对模型的性能影响**

本文使用GPT-4对不包含同义词信息的术语产生3个同义词，在对同义词的使用数量进行探索的过程中分别尝试添加1个，2个，以及3个同义词，基于三种基座模型BioBERT、Bioformer和PubMeBERT在GSC+数据集上进行性能对比。实验结果如图6所示。

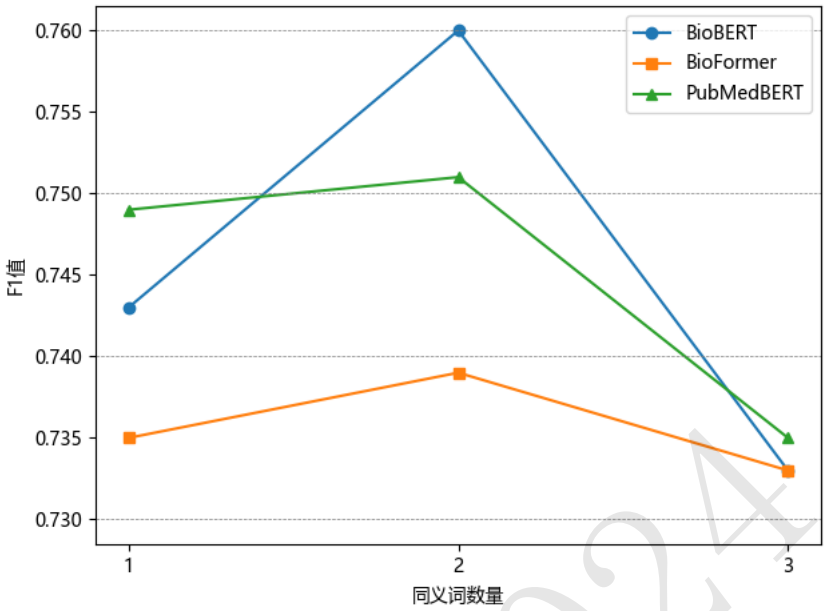



Figure 6: 同义词数量在GSC+数据集上对模型的性能影响

实验结果表明：在三种基座模型上，均为添加2个同义词时获得了最佳的性能。其原因可能是GPT-4生成2个同义词最为准确，因此添加2个同义词可以学习到最丰富的语义信息，而添加1个和3个同义词时出现了信息不足和干扰信息过多的情况。因此后续实验均选择使用2个同义词。

**C.本文方法详细结果**

本文方法可以在NER和NEN子任务上分别进行评价，本文最后使用的GPT+EMB融合后模型在GSC+和ID-68数据集两个数据集上的NER和NEN任务的效果如表9所示。

| 子任务 | GSC+数据集 | | | ID-68数据集 | | |
|---|---|---|---|---|---|---|
| | Precision | Recall | F1 | Precision | Recall | F1 |
| NER | 0.941 | 0.848 | 0.880 | 0.964 | 0.837 | 0.888 |
| NEN | 0.783 | 0.744 | 0.763 | 0.924 | 0.781 | 0.846 |

Table 9: 本文方法详细结果

第二十三届中国计算语言学大会论文集，第553页-第567页，太原，中国，2024年7月25日至28日。
卷1：主会论文
(c) 2024 中国中文信息学会计算语言学专业委员会

# 基于机器学习的语音情感声学特征筛选

**董文琪**[1]，**王涵**[1]，**张璟玮**[12*]
[1]澳门大学人文学院中国语言文学系
[2]澳门大学认知与脑科学研究中心
yc37717@um.edu.mo；yc37729@um.edu.mo；jwzhang@um.edu.mo

## 摘要

筛选有效表达情感的声学特征对语音情感研究至关重要。对具有相同或相似声学特征的情感，声学研究中仅使用基频和时长无法有效区分。本研究扩大声学参数的种类和数量，使用三种机器学习方法，筛选出区分情感类型的多组有效声学参数，补充和完善语音情感声学研究的声学特征集。研究发现，区分不同情感所依赖的声学参数、参数数量、参数贡献都不相同，其中频谱和信噪参数发挥重要作用。本研究为语音情感声学分析的参数选择提供参考。

**关键词：** 声学特征；语音情感分类；机器学习；频谱；噪音水平

# Acoustic Feature Selection for Speech Emotion Based on Machine Learning

**DONG Wenqi**[1], **WANG Han**[1], **ZHANG Jingwei**[12*]
[1]Department of Chinese Language and Literature, Faculty of Arts and Humanities, University of Macau
[2]Centre for Cognitive and Brian Sciences, University of Macau
yc37717@um.edu.mo;yc37729@um.edu.mo;jwzhang@um.edu.mo

## Abstract

The selection of acoustic features that effectively convey emotions is crucial for research on speech emotion. For emotions with similar or identical acoustic characteristics, the differentiation solely through fundamental frequency and duration proves inadequate within acoustic studies. This study broadens the types and quantities of acoustic parameters, employing three machine learning methods to identify multiple sets of effective acoustic parameters for distinguishing emotion types. These augment and refine the acoustic feature set for speech emotion research. Findings reveal that the acoustic parameters, their quantities, and contributions vary in discriminating different emotions, with spectral and signal-to-noise parameters playing pivotal roles. This study offers insights into parameter selection for acoustic analysis of speech emotion.

**Keywords:** Acoustic features , Speech emotion classification , Machine learning , Spectrum , Noise level

通讯作者
基金项目：澳门特别行政区科学技术发展基金0036/2022/ITP
©2024中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

## 1 引言

语音能够传达说话者的非语言信息，包括生理特征（例如身高、体型、性别、年龄）、社会特征（例如社会阶层、受教育程度）和情感特征（例如愤怒、高兴、悲伤、恐惧）（Laver and Trudgill，1979）。不同领域对语音情感研究的侧重不同，例如计算机科学主要关注语音情感识别的算法和情感识别的准确性，声学语音学聚焦于不同情感状态下语音的声学特征。不论是研究语音情感识别还是语音情感的声学特征，都需要基于一组能有效区分不同情感类型的声学参数。

声学参数表示语音信号在频率、振幅、响度等方面的物理特性（Bhangale and Kothandaraman，2023），是声学特征的量化指标。早期的语音学研究发现，基频是区分不同情感最常用的声学参数之一（Cosmides，1983）。例如，愤怒和高兴的平均基频更高，基频变化范围更大（Pereira and Watson，1998；Yildirim et al.，2004）。与时长、能量相关的参数也是常用的声学参数。愤怒和高兴通常持续时长更长（Yildirim et al.，2004），恐惧则表现为持续时长缩短、语速变快（Kienast and Sendlmeier，2000）。Rong等（2009）的研究从84个特征中筛选出了16个最有效的声学特征，其中大多数与基频和能量相关。有的情感类型在基频等声学特征上有相同或相似的表现，例如高兴和愤怒（Pereira and Watson，1998；Zvarevashe and Olugbara，2020），因此仅通过基本的声学参数不能实现语音情感的良好区分。在语音情感识别中最常用的声学特征包括时域的韵律特征，例如基频、时长、能量，和频域的谱特征，例如线性预测倒谱系数（LPCC）、梅尔频率倒谱系数（MFCC）（Zvarevashe and Olugbara，2020；刘振焘等，2018；陶建华等，2023）。近年来，各领域的研究者也越来越注意到声门特征对语音情感识别的重要性（Sun and Zhang，2018）。表示声门特征的参数，例如基频抖动（Jitter）、振幅微扰（Shimmer）应用到了语音情感识别中（陶建华等，2023）。在语音学研究中，Birkholz等（2015）通过感知实验发现，嗓音发声类型对离散情感的分类起到重要作用，其中中性情感和快乐表现为模态发声（modal voice），恐惧表现出明显的气嗓音（breathy voice），愤怒表现为较紧的嗓音（pressed voice）。张锐锋（2016）从声门阻抗信号（EGG）中提取了Jitter、Shimmer和谐波噪音比（Harmonic Noise Ratio，HNR）参数，有效区分了高兴和愤怒两种情感。

近年来，许多语音情感识别的研究都聚焦于分类器选择和算法优化上，而对于许多机器学习算法来说，语音声学参数的选择和精简更为重要（徐欣等，2014；Sun，2020；Jha et al.，2022）。目前的研究已经得出从语音信号中提取的声学特征是语音情感识别的重要依据，但还不清楚哪些特征对语音情感识别更有效（Sun，2020；Zvarevashe and Olugbara，2020）。对于情感声学特征的研究，大多是比较不同情感在某一个或几个常用声学参数上的异同（董理等，2021），对情感的声学特征分析还不够系统全面。本研究将基于实验控制下录制的语音样本形成小型数据集，通过监督学习的方法，筛选出能够区分不同情感并且具有语言学意义的声学参数，补充完善语音情感声学研究的特征集，进而促进情感语音合成的研究。

## 2 实验方法与设计

### 2.1 录音材料

词是语音情感识别最小的有意义的单位(Schuller et al.，2011)。同时，不同音素的特征对情感识别也有影响(Sethu et al.，2008)。考虑到情感表达的完整性和元音对情感识别的影响，本研究选择“dada”这一不预设任何字形、字义和声调的无意义音节作为录音材料。录音时要求发音人分别以中性、高兴、愤怒和恐惧四种情感状态读出这个词。

### 2.2 发音人

共有21人进行了录音，11名男性10名女性，平均年龄为25.76±4.52岁。

### 2.3 录音设备及条件

录音地点是一间专业的隔音室，使用便携式录音笔(奥林巴斯LS-100)和电容式麦克风(AKG C-420)录制语音信号，采样频率为44,100 Hz，精度值为16 bit，保存为.wav格式，最后将录制好的语音导入联想Thinkpad T14型笔记本电脑进行预处理。录音时每个例词读两遍，两遍中间间隔时长为1秒。


(c) 2024 中国中文信息学会计算语言学专业委员会

### 2.4 数据预处理

首先，使用Praat对录制的语音样本进行切音和标注处理。切音时以双音节词为单位切分，标注时对每个音节都选主要元音段进行标注。为了使数据标注具有内部一致性，语音波形起点选择在声带规则振动的第二个周期的起始点，终点选择在基频曲线、振幅曲线、第一共振峰和第二共振峰结构相对稳定的点。经过预处理后，共得到336个样本(4种情感*2遍*2个音节*21人)，每种情感类型各有84个样本。

其次，通过基于Matlab的语音声学分析软件VoiceSauce提取声学参数。VoiceSauce是针对语音学研究而开发的声学参数测量、提取与分析软件(Shue et al.，2011；凌锋等，2019)。该软件包含33个与语音学相关的常用声学参数，可以分为以下四类：

(1)基本参数：基频(F0)、第一到第四共振峰(F1/F2/F3/F4)、能量(Energy)，从时域信号中提取的参数，是语音情感研究最基本、最常用的声学参数。其中，基频参数的测量常用的有Stright(strF0)和Praat(pF0)两种算法，共振峰参数的测量常用的有Snack(sF1/sF2/sF3/sF4)和Praat(pF1/pF2/ pF3/ pF4)两种算法。

(2)频谱参数：第一谐波(H1)、第二谐波(H2)、第四谐波(H4)、靠近2000 Hz的谐波(H2K)、靠近第一共振峰的谐波(A1)、靠近第二共振峰的谐波(A2)、靠近第三共振峰的谐波(A3)、第一、二谐波差(H1-H2)、第二、四谐波差(H2-H4)、第四与靠近2000 Hz的谐波差(H4-H2K)、靠近2000 Hz的谐波与靠近5000 Hz的谐波差(H2K- H5K)、第一谐波与第一共振峰谐波振幅差(H1-A1)、第一谐波与第二共振峰谐波差(H1-A2)、第一谐波与第三共振峰谐波差(H1-A3)、次谐波与谐波比(SHR)。从频域信号中提取的参数，反映的是声带闭合程度和声带闭合速率，进而反映语音信号在不同频段上的能量变化速度及嗓音发声类型(Stevens, 1977)。一般来说，谐波振幅差值越大，频谱斜率越大，声带闭合程度越低，闭合速率低，能量下降速度越快，表现为气嗓音。

(3)信噪参数：倒频谱突显峰值(CPP)、0-500 Hz频段的谐波噪音比(HNR05)、0-1500 Hz频段的谐波噪音比(HNR15)、0-2500 Hz频段的谐波噪音比(HNR25)、0-3500 Hz频段的谐波噪音比(HNR35)，反映语音中噪音成分的多少和信号周期性的好坏。CPP和HNR的值越大，语音中的噪音成分越少，信号的周期性越好。

(4)声源激励参数：激励期(epoch)、激励强度(SoE)，反映声门状态。激励期即声门闭合瞬间，激励强度受到声门闭合时长和闭合速率的影响。

### 2.5 算法选择与评估

每一种算法都有其优缺点。考虑到数据量的大小、分类的多少以及算法的难易程度，本研究选择K最近邻(K-Nearest Neighbors，以下简称“KNN”)(Petrushin，2000)、支持向量机(Support Vector Machine，以下简称“SVM”)(Schuller et al.，2004)和随机森林(Random Forest，以下简称“RF”)(Breiman，2001)三种经典机器学习算法对语音情感的声学特征进行筛选。KNN是最基本、最简单的分类算法，对语音情感数据的拟合性能较高(Basu et al.，2017；Jha et al.，2022)，SVM的二分类效果最佳，并且适合小样本训练集(Jha et al.，2022)，RF是基于决策树的集成学习算法，能够避免由较小数据集引起的过度拟合问题(Rong et al.，2009；Jha et al.，2022)。为了便于比较三种算法的分类效果，本研究将中性、高兴、愤怒、恐惧四种情感两两组合，均进行二分类。每种分类模型均按照70%的比例设置训练集，30%的样本作为测试集。采用准确率(Accuracy)这一最直观的衡量标准来比较各算法在不同情感组合中的分类准确性，它是预测正确的样本数与样本总数的比率(Zvarevashe and Olugbara，2020)。通过受试者工作特征曲线(Receiver Operating Characteristic Curve，以下简称“ROC曲线”)和ROC曲线下面积(Area Under Curve，以下简称“AUC”)来评估各算法的分类性能，ROC曲线和AUC是评估二分类模型性能的重要指标(Jha et al.，2022)。

## 3 结果

### 3.1 分类效果

在设置模型参数时，均通过定义一个超参数网格并通过网格搜索来寻找最佳的参数组合，以提高模型的性能。准确率也是在最佳参数组合下得出的。

从表1中可以看出，不同的算法在区分不同情感时的准确率不同。平均来看，三种算法区分愤怒和恐惧、高兴和恐惧的效果最好，平均准确率均超过85%；区分中性与恐惧的效果最


(c) 2024 中国中文信息学会计算语言学专业委员会

差，平均准确率只有64%。具体的，KNN和SVM区分愤怒和恐惧效果最好，KNN的准确率达到了96%，SVM的准确率达到92%；RF区分高兴和恐惧的准确率最高，为84%。其次，在三种算法中，相比较而言，高兴情感都能与其他三种情感进行良好区分，而且高兴与恐惧之间区分得最好，在KNN中准确率达到88%，SVM的准确率为84%，RF的准确率也是84%。不论哪一种算法，恐惧与中性情感的区分效果都比较差，在KNN和RF中，区分恐惧与中性情感的准确率都是最低的，只有60%；在SVM中，区分恐惧与中性情感的准确率为72%，只高于愤怒与中性情感的区分准确率。此外，在KNN和SVM中，愤怒与中性情感的区分效果也比较差，在KNN中的准确率最低，仅有64%。总之，在三种算法中，高兴、愤怒和恐惧三种情感之间能够进行较为准确的区分，恐惧和中性情感最难准确区分开。

| 情感类型 | KNN | SVM | RF | 平均准确率 |
|---|---|---|---|---|
| 中性/高兴 | 88% | 76% | 72% | 78.67% |
| 中性/愤怒 | 64% | 68% | 80% | 70.67% |
| 中性/恐惧 | 60% | 72% | 60% | 64% |
| 高兴/愤怒 | 84% | 76% | 80% | 80% |
| 高兴/恐惧 | 88% | 84% | 84% | 85.33% |
| 愤怒/恐惧 | 96% | 92% | 72% | 86.67% |

Table 1: 基于三种算法二分类的准确率

对中性/高兴、中性/愤怒、中性/恐惧、高兴/愤怒、高兴/恐惧、愤怒/恐惧六组情感都使用KNN、SVM和RF三种算法进行二分类，图1—图6是对基于三种算法的分类模型性能的评估结果，用ROC曲线表示。ROC曲线横轴为假正例率(False Positive Rate，以下简称“FPR”)，纵轴为真正例率，也称为召回率(True Positive Rate，以下简称“TPR”)。FPR表示将负例误判为正例的比例，TPR表示将正例正确判断为正例的比例。FPR越低，TPR越高，ROC曲线越靠近左上角，说明模型的性能越好。

通过图1可以看出，三种分类模型在区分中性-高兴时的FPR都为0，说明对两种情感分类时没有误判，KNN的TPR要高于SVM和RF的TPR，说明在区分中性-高兴时，KNN算法更灵敏，召回率更高。从图2中可以看出，相比较而言，KNN在区分中性-愤怒时的FPR较高，TPR较低，KNN算法区别中性-愤怒两类情感时效果较差。图3显示，三种分类模型在区分中性-恐惧时的FPR都较低，即对两种情感的误判较低，但TPR也很低，说明对中性和恐惧这两类情感的分类不灵敏。对比图1—图3可以发现，KNN、SVM和RF三种算法区分对中性与愤怒、中性与恐惧的区分效果不佳。

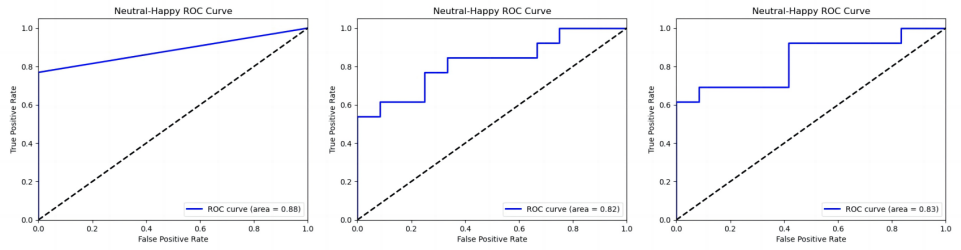



Figure 1: KNN、SVM、RF区分“中性/高兴”的ROC曲线


(c) 2024 中国中文信息学会计算语言学专业委员会

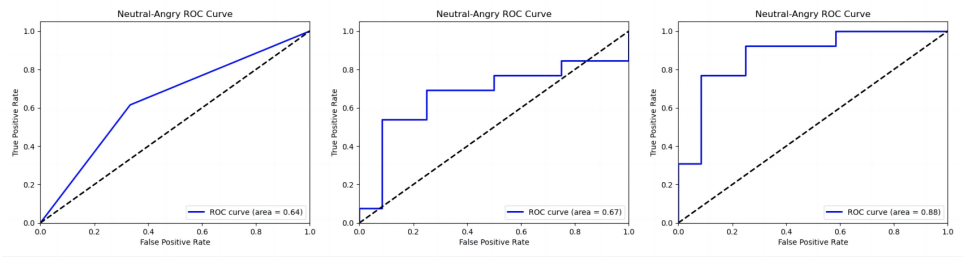



Figure 2: KNN、SVM、RF区分“中性/愤怒”的ROC曲线

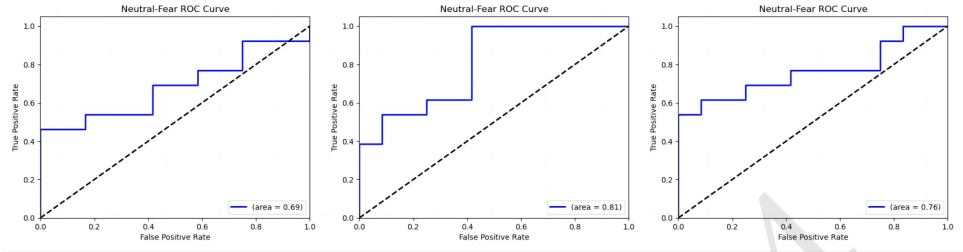



Figure 3: KNN、SVM、RF区分“中性/恐惧”的ROC曲线

通过图4—图6可以看出，相比较而言，不论是区分高兴-愤怒、高兴-恐惧还是愤怒-恐惧，基于KNN算法的分类模型都存在将两种情感误判情况。对高兴-愤怒的区分中，基于RF算法的模型分类效果最好，对高兴-恐惧的区分也是如此。区分愤怒-恐惧时，基于三种算法的分类模型效果都很好。

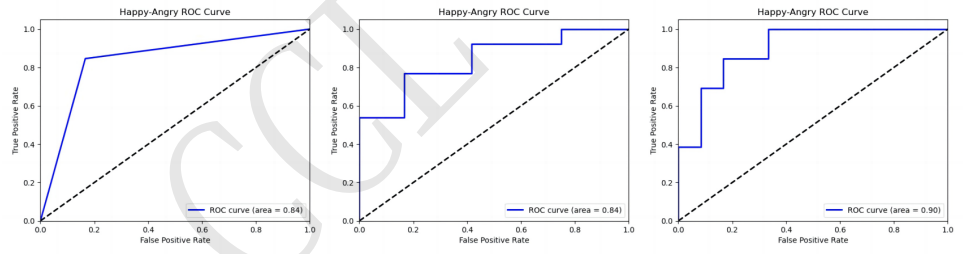



Figure 4: KNN、SVM、RF区分“高兴/愤怒”的ROC曲线

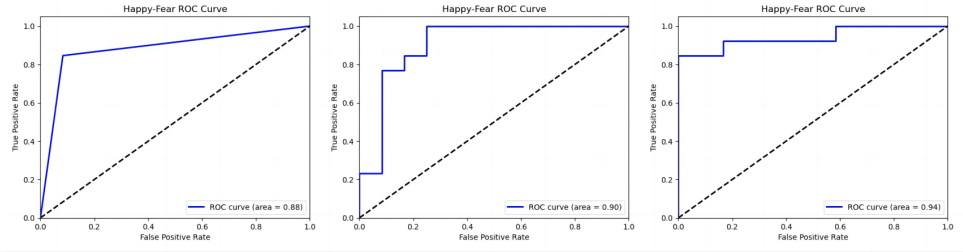



Figure 5: KNN、SVM、RF区分“高兴/恐惧”的ROC曲线


(c) 2024 中国中文信息学会计算语言学专业委员会

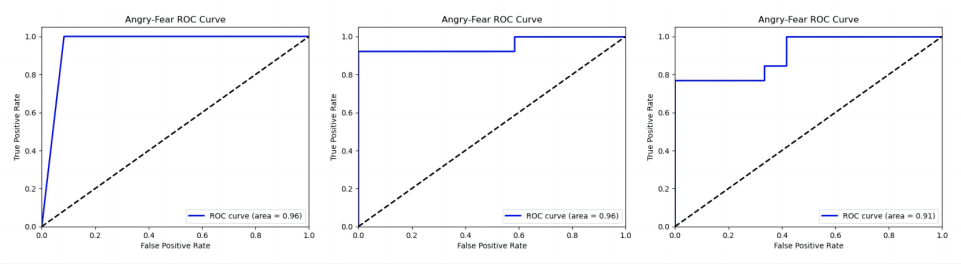



Figure 6: KNN、SVM、RF区分“愤怒/恐惧”的ROC曲线

### 3.2 特征筛选

基于不同算法的不同情感之间的分类都使用了相同的一组参数。在进行模型训练之前，首先通过独立样本T检验初步筛选出与情感分类不存在显著性差异(ps $<0.05$)的参数并剔除，这部分参数如附录A所示。

基于剩下的声学特征进行模型训练与测试。排列重要性(Permutation Importance)是衡量各个特征对模型性能影响程度的指标。通过这一方法可以得出不同声学参数对模型的贡献，进而筛选出帮助分类的最有效的一组声学参数。排列重要性是基于特定模型的，本研究中基于KNN、SVM和RF三种算法得出的各组声学参数的排序均不相同。我们已经得出评估不同模型的AUC(表2)，因此本研究提出了一种综合三种算法的排序的计算方法，对各声学参数在不同情感分类中的贡献进行综合计算和排序。计算公式如下：

$$\text{Weight} = k_{\text{KNN}} \cdot s_{\text{KNN}} + k_{\text{SVM}} \cdot s_{\text{SVM}} + k_{\text{RF}} \cdot s_{\text{RF}}$$

其中，“Weight”代表声学参数在排序中所占的权重，即贡献。“k”表示各算法占比系数，k值根据AUC计算，即对每一组情感，以三种算法中最小的AUC值(即模型效果最差)为基线，对三个AUC值进行归一化，归一化后的值即为k。“s”表示不同算法得出的各声学参数排序重要性的得分。声学参数的最终贡献为三个算法的占比系数与该算法下声学参数的重要性得分的乘积之和。

| 情感类型 | KNN | SVM | RF |
|---|---|---|---|
| 中性/高兴 | 88% | 82% | 83% |
| 中性/愤怒 | 64% | 67% | 88% |
| 中性/恐惧 | 69% | 81% | 76% |
| 高兴/愤怒 | 84% | 84% | 89% |
| 高兴/恐惧 | 88% | 90% | 94% |
| 愤怒/恐惧 | 96% | 96% | 91% |

Table 2: 评估三种算法二分类的AUC

基于上述计算方法，我们得出区分各组情感的声学参数重要性排序及其贡献，如表3所示。对于部分参数贡献出现负值的情况，根据Scikit-learn的解释，这可能是由于使用的数据集较小或者参数之间存在较强相关性导致的，未来可以通过聚类分析进行更细致的参数筛选改善(Pedregosa et al.，2011)。

从表3中可以看出，已有语音情感声学研究常用的基本声学参数基频(strF0和pF0)和时长(Time)对不同情感的区分没有贡献或贡献较低。基频在中性/高兴和高兴/恐惧两组情感分类中有贡献，时长对中性/恐惧的区分有贡献。在中性/高兴组中，贡献最大的H1-A2得分是基频的11倍，在高兴/恐惧组中，相比较而言，基于Stright算法得出的基频(strF0)贡献更大，但贡献最大的CPP得分约是它的1.6倍。在中性/恐惧组中，贡献最大的H4的得分约是时长的5.4倍。

其次，区分不同情感所依赖的声学参数、参数数量、参数的贡献都不相同。在区别中性/高兴、中性/愤怒时H1-A2的贡献最大，区别中性/恐惧时H4的贡献最大，区别高兴/愤怒时SHR的贡献最大，区别高兴/恐惧、愤怒/恐惧时CPP的贡献最大。H1-A2，即第一谐波与第二共振峰谐波振幅差，反映声带闭合的速率(Stevens，1977)，进而反映声带漏气情况和能量下


(c) 2024 中国中文信息学会计算语言学专业委员会

降的快慢。H4指频谱上的第四谐波，其频率是基频的四倍。在语音学研究中，谐波成分很少单独使用，常通过进一步计算谐波差来反映语音信号的频谱斜率(Gordon and Ladefoged，2001)，它反映声带的闭合情况，进而反映频谱能量的变化。SHR是次谐波与谐波比，影响语音信号非周期性振动段的音高，用来判断信号的周期性(凌锋等，2019)。CPP即倒频谱突显峰值，测量的是倒谱峰值到倒谱回归线的距离，反映语音信号中噪音成分的多少，进而反映语音信号周期性的好坏(Shue et al.，2011；凌锋等，2019)。这四个参数及其相关参数可以分为频谱参数和信噪参数两类，都是判断嗓音发声类型的常用参数，说明嗓音发声类型对情感识别发挥了重要作用。在语音情感的声学分析中，除了常用的基频、时长、能量等参数，还需要引入和分析反映嗓音发声类型的频谱参数和信噪参数。

对语音学研究来说，表3为语音情感的声学分析提供了更重要、更全面的声学特征参考。对中性、高兴、愤怒、恐惧四种情感进行声学分析时，可根据研究者的需要，选择贡献更大、排序更靠前的参数进行分析和声学特征比较。

| 排序 | 中性/高兴 | | 中性/愤怒 | | 中性/恐惧 | | 高兴/愤怒 | | 高兴/恐惧 | | 愤怒/恐惧 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 特征 | 贡献 | 特征 | 贡献 | 特征 | 贡献 | 特征 | 贡献 | 特征 | 贡献 | 特征 | 贡献 |
| 1 | H1-A2 | 0.22 | H1-A2 | 0.30 | H4 | 0.49 | SHR | 0.27 | CPP | 0.21 | CPP | 0.17 |
| 2 | SHR | 0.18 | H2K | 0.26 | Energy | 0.22 | pF1 | 0.15 | HNR05 | 0.18 | A1 | 0.08 |
| 3 | SoE | 0.14 | pF1 | 0.24 | A2 | 0.18 | H2K | 0.11 | strF0 | 0.13 | Energy | 0.08 |
| 4 | pF3 | 0.11 | Energy | 0.22 | H1-A1 | 0.13 | A1 | 0.03 | HNR15 | 0.09 | H2 | 0.08 |
| 5 | H2 | 0.09 | H4 | 0.22 | HNR15 | 0.12 | Energy | -0.02 | A1 | 0.09 | HNR05 | 0.07 |
| 6 | H1-A3 | 0.08 | A2 | 0.19 | A1 | 0.10 | HNR35 | -0.02 | A3 | 0.06 | H1-H2 | 0.06 |
| 7 | H1-A1 | 0.06 | HNR15 | 0.18 | Time | 0.09 | HNR15 | -0.02 | pF0 | 0.06 | pF1 | 0.06 |
| 8 | pF0 | 0.02 | HNR25 | 0.15 | H2K | 0.06 | H4 | -0.03 | pF2 | 0.06 | H1-A2 | 0.03 |
| 9 | strF0 | 0.02 | A1 | 0.13 | A3 | 0.06 | H2 | -0.03 | epoch | 0.04 | A2 | 0.02 |
| 10 | H2-H4 | 0.02 | A3 | 0.08 | H1-A2 | 0.05 | HNR25 | -0.04 | H2K | 0.04 | pF3 | 0.02 |
| 11 | H1 | 0.00 | H1-A1 | 0.08 | H1-A3 | 0.02 | HNR05 | -0.04 | HNR25 | 0.03 | H1-A3 | 0.02 |
| 12 | pF4 | -0.02 | HNR35 | 0.08 | | | SoE | -0.06 | pF3 | 0.03 | SHR | 0.02 |
| 13 | HNR05 | -0.07 | H1-A3 | 0.01 | | | H1 | -0.06 | sF2 | 0.03 | sF2 | 0.02 |
| 14 | epoch | -0.07 | | | | | A3 | -0.06 | H1-A2 | 0.02 | H4-H2K | 0.02 |
| 15 | | | | | | | A2 | -0.10 | HNR35 | 0.02 | H4 | 0.00 |
| 16 | | | | | | | epoch | -0.11 | A2 | 0.02 | A3 | 0.00 |
| 17 | | | | | | | strF0 | -0.15 | H1-A3 | 0.01 | pF2 | 0.00 |
| 18 | | | | | | | pF0 | -0.18 | H4-H2K | 0.0 | H2K | -0.02 |
| 19 | | | | | | | | | H1-A1 | 0.00 | H1-A1 | -0.04 |
| 20 | | | | | | | | | SoE | 0.00 | | |
| 21 | | | | | | | | | H2-H4 | -0.02 | | |
| 22 | | | | | | | | | Energy | -0.04 | | |

Table 3: 声学参数重要性排序及贡献(保留到小数点后两位)

## 4 结论

在语音情感研究中，不同领域的研究侧重点不同，语音学更关注语音中不同情感的声学表现。现有的语音情感的声学研究大多是探讨两到三种离散的情感在基频、时长等一个或少数几个常用参数上的不同表现，而有些情感在这些常用参数上存在许多共同特征，例如愤怒和高兴的平均基频相较于中性情感都变大，因此缺乏区别性的声学特征，筛选有效表达情感的声学特征至关重要。本研究通过扩大声学参数的种类和数量，使用机器学习的方法，筛选出区分情感类型的一组有效声学参数，以补充和完善语音情感声学研究的声学特征集。

研究采用机器学习的方法，将语音学研究常用的四类共33个声学参数放入基于KNN、SVM和RF三种算法的分类模型中，筛选出不同情感分类所依赖的多组声学参数。结果发现，语音情感声学分析常用的基频和时长参数对各组情感的分类贡献不大。区分不同情感所依赖的声学参数、参数数量、参数的贡献都不相同。H1-A2在区分中性/高兴、中性/愤怒时的贡献最大，H4在区分中性/恐惧时的贡献最大，SHR在区分高兴/愤怒时的贡献最大，CPP在区分高兴/恐惧、愤怒/恐惧时的贡献最大。这组参数表明频谱参数和信噪参数对区分不同情感类型具有重要作用。本研究得出的声学参数排序重要性为语音情感声学分析的参数选择提供参考。

第二十三届中国计算语言学大会论文集，第568页-第576页，太原，中国，2024年7月25日至28日。
卷1：主会论文
(c) 2024 中国中文信息学会计算语言学专业委员会

由于本研究所使用的数据集较小，在算法选择等方面受到限制，后续研究应在此基础上扩大样本量，以优化算法，从而获得更精简和精确的声学参数。此外，本研究通过数据归一化来消除性别等因素上的差异，已有研究发现性别影响不同情感的区分(Arias et al.，2021)，未来可以在获得大量数据的基础上，探讨性别对各组情感分类的影响。在语音学研究方面，本研究为语音情感声学分析选择哪些参数提供了参考，后续可以在此基础上进行声学分析，探讨不同情感在这些参数上的具体表现。

## 参考文献


Arias P. Rachman L. Liuni M. and Aucouturier J.-J. 2021. Beyond Correlation: Acoustic Transformation Methods for the Experimental Study of Emotional Voice and Speech. *Emotion Review*, 13(1), 12–24.

Bhangale K. and Kothandaraman M. 2023. Speech Emotion Recognition Based on Multiple Acoustic Features and Deep Convolutional Neural Network. *Electronics*, 12(4), Article 4.

Birkholz P. Martin L. Willmes K. Kröger B. J. and Neuschaefer-Rube C. 2015. The contribution of phonation type to the perception of vocal emotions in German: An articulatory synthesis study. *The Journal of the Acoustical Society of America*, 137(3), 1503–1512.

Breiman L. 2001. Random Forests. *Machine Learning*, 45(1), 5–32.

Cosmides L. 1983. Invariances in the acoustic expression of emotion during speech. *Journal of Experimental Psychology: Human Perception and Performance*, 9(6), 864–881.

Arias P. Rachman L. Liuni M. and Aucouturier J.-J. 2021. Beyond Correlation: Acoustic Transformation Methods for the Experimental Study of Emotional Voice and Speech. *Emotion Review*, 13(1), 12–24.

Gordon M. and Gordon P. 2001. Phonation types: a cross-linguistic overview. *Journal of phonetics*, 29(4), 383-406.

Jha T. Kavya R. Christopher J. and Arunachalam V. 2022. Machine learning techniques for speech emotion recognition using paralinguistic acoustic features. *International Journal of Speech Technology*, 25(3), 707–725.

Kienast M. and Sendlmeier W. F. 2000. Acoustical Analysis of Spectral and Temporal Changes in Emotional Speech. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion.*

Laver J. and Trudgill P. 1979. Phonetic and linguistic markers in speech. In: Scherer, Klaus R.; Giles, Howard (Ed.). *Social markers in speech.* Cambridge: Cambridge University Press.p.1-32.

Pedregosa et al. 2011. Scikit-learn: Machine Learning in Python. *JMLR 12*, pp. 2825-2830.

Pereira C. and Watson C. 1998. Some acoustic characteristics of emotion. *5th International Conference on Spoken Language Processing (ICSLP 1998)*, paper 0684-0.

Petrushin V. A. 2000. Emotion recognition in speech signal: Experimental study, development, and application. *6th International Conference on Spoken Language Processing (ICSLP 2000)*, vol.2, 222-225–0.

Rong J. Li G. and Chen Y.-P. P. 2009. Acoustic feature selection for automatic emotion recognition from speech. *Information Processing  Management*,45(3), 315–328.

Schuller B. Rigoll G. and Lang M. 2004. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*,1, I–577.

Schuller B. Batliner A. Steidl S. Seppi D. 2011. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication*, 53(9), 1062–1087.

Sethu V. Ambikairajah E. and Epps J. 2008. Phonetic and speaker variations in automatic emotion classification. *Interspeech 2008* , (pp.617-620).

Shue Y. L. Keating P. Vicenik C. and Yu K. 2011. VoiceSauce: A program for voice analysis. *ICPhS XVII* , (pp.1846-1849).



(c) 2024 中国中文信息学会计算语言学专业委员会

Stevens K. N. 1977. Physics of laryngeal behavior and larynx modes. *Phonetica*, 34(4), 264-279.

Sun T.-W. 2020. End-to-End Speech Emotion Recognition With Gender Information. *IEEE Access*, 8, 152423–152438.

Sun Y. and Zhang X-Y. 2018. Characteristics of human auditory model based on compensation of glottal features in speech emotion recognition. *Future Generation Computer Systems*, 81, 291–296.

Yildirim S. Bulut M. Lee C. M. Kazemzadeh A. Deng Z. Lee S. Narayanan S. and Busso C. 2004. An acoustic study of emotions expressed in speech. *Interspeech 2004*,2193–2196.

Zvarevashe K. and Olugbara O. 2020. Ensemble Learning of Hybrid Acoustic Features for Speech Emotion Recognition. *Algorithms*, 13(3), Article 3.

董理, 梁晓静, 黄慧怡. 2021. 昆曲女性行当情感念白时长特征. 语言学论丛(02),272-290.

凌锋, 史辉, 袁丹, 沈瑞清. 2019. 发声态研究的相关问题与VoiceSauce的使用. 方言(04),385-397.

刘振焘, 徐建平, 吴敏, 曹卫华, 陈略峰, 丁学文, 郝曼, 谢桥. 2018. 语音情感特征提取及其降维方法综述. 计算机学报(12),2833-2851.

陶建华, 陈俊杰, 李永伟. 2023. 语音情感识别综述. 信号处理(04),571-587.

徐欣, 李雅, 许小颖, 陶建华. 2014. 情感语音识别的区别性声学特征选择.第十一届中国语音学学术会议(PCC2014)论文集(pp.146).

张锐锋. 2016. 普通话情感语句中的发声——一项预试性研究. 语言学论丛(02),305-322.

## 附录A. 初步剔除的参数

| 情感类型 | 剔除的参数 |
|---|---|
| 中性/高兴 | Time、H4、H2K、A1、A2、A3、H1-H2、H4-H2K、H2K-H5K、CPP、Energy、HNR15、HNR25、HNR35、sF1、sF2、sF3、sF4、pF1、pF2 |
| 中性/愤怒 | Time、H1、H2、H1-H2、H2-H4、H4-H2K、H2K-H5K、CPP、HNR05、SHR、strF0、pF0、sF1、sF2、sF3、sF4、pF2、pF3、pF4、epoch、SoE |
| 中性/恐惧 | H1、H2、H4、H2K、A1、A2、A3、H1-H2、H2-H4、H4-H2K、H2K-H5K、H1-A1、H1-A2、H1-A3、strF0、pF0、sF1、sF3、sF4、pF1、pF4、epoch |
| 高兴/愤怒 | Time、H1-H2、H2-H4、H4-H2K、H2K-H5K、H1-A1、H1-A2、H1-A3、CPP、sF1、sF2、sF3、sF4、pF2、pF3、pF4 |
| 高兴/恐惧 | Time、H1、H2、H4、H1-H2、H2K-H5K、SHR、sF1、sF3、sF4、pF1、pF4 |
| 愤怒/恐惧 | Time、H1、H2-H4、H2K-H5K、HNR15、HNR25、HNR35、strF0、pF0、sF1、sF3、sF4、pF4、epoch、SoE |

CCL 2024

# 基于交互行为语义模式增强的ID推荐方法

**王远来**[1], **白宇**[*1,2], **廉鹏**[2,3]
[1]沈阳航空航天大学计算机学院，辽宁沈阳，110136
[2]多语言协同翻译技术国家地方联合工程实验室，辽宁沈阳，110136
[3]沈阳北软信息职业技术学院，辽宁沈阳，110136
wangyuanlai@stu.sau.edu.cn; baiyu@sau.edu.cn; lianp@ge-soft.com

## 摘要

基于ID的推荐是一种依赖用户或物品的唯一标识符进行推荐的经典推荐方法，这种方法经常面临用户物品交互数据稀疏、符号ID缺失语义信息等问题。该文针对上述问题，假设不同领域的用户-物品交互行为之间存在潜在的模式关联，提出了一种基于交互行为语义模式增强的ID推荐方法。该方法在目标域推荐任务中引入辅助域信息，基于图神经网络对辅助域和目标域信息进行联合编码表示，通过引入交互行为语义模式，将辅助域的用户-物品交互信息以及物品描述信息迁移至目标域，从而实现目标域ID推荐中的交互行为语义增强。在8个公开数据集上的实验结果表明，相比目前的SOTA模型，本文方法表现出更好的推荐效果，其Recall@20与NDCG@20分别具有3% ∼ 30%、1% ∼ 40%的提升。

**关键词:** 推荐系统；交互行为；语义模式；语义增强

# Enhanced ID Recommendation Method Utilizing Semantic Patterns of Interactive Behaviors

**Yuanlai Wang**[1], **Yu Bai**[*1,2], **Peng Lian**[2,3]
[1]School of Computer Science, Shenyang Aerospace University, Shenyang, Liaoning, 110136
[2]National & Local Joint Engineering Laboratory of Multi-Language Collaborative Translation Technology, Shenyang, Liaoning, 110136
[3]Shenyang Northern Software College of Information Technology, Shenyang, Liaoning, 110136
wangyuanlai@stu.sau.edu.cn; baiyu@sau.edu.cn; lianp@ge-soft.com

## Abstract

ID-based recommendation is a traditional method in recommendation systems that relies on unique identifiers of users or items for recommendation. However, it often faces challenges such as sparse user-item interaction data and a lack of semantic information in symbolic IDs. This study addresses these issues by suggesting the existence of latent pattern correlations between user-item interaction behaviors across different domains. It proposes an improved ID-based recommendation method that incorporates semantic patterns of interaction behaviors. This approach integrates auxiliary domain information into the target domain recommendation task by utilizing graph neural networks to jointly encode the auxiliary and target domain information. By including semantic patterns of interaction behaviors, it facilitates the transfer of user-item interaction information and item description data from the auxiliary domain to the target domain, thereby enhancing the semantic aspects of interaction behavior in ID-based recommendation for the target domain. Experimental results on eight public datasets have demonstrated that, in comparison to current state-of-the-art (SOTA) models, the

* 通讯作者 Corresponding author.

第二十三届中国计算语言学大会论文集，第577页-第587页，太原，中国，2024年7月25日至28日。
卷1：主会论文
(c) 2024 中国中文信息学会计算语言学专业委员会

proposed method achieves superior recommendation performance, with enhancements in Recall@20 ranging from 3% to 30% and NDCG@20 ranging from 1% to 40%.

**Keywords:** Recommended system , Interactive behavior , Semantic pattern , Semantic enhancement

## 1 引言

推荐系统(Wang et al., 2021)因其能向用户提供个性化推荐服务而具有广泛的应用场景(Yu et al., 2022; Huang et al., 2021)。对推荐系统的研究一直是工业界和学术界关注的焦点之一。

基于ID的推荐是一种简单直接的推荐系统实现方法，其主要利用用户和物品的交互信息实现个性化推荐。通常，推荐系统将用户与物品视为基本处理单元，并为其分配唯一的离散标识(ID)，用户和物品的交互信息通常被表示为User-ID与Item-ID的对应关系集合。基于ID的推荐系统通过对这些对应关系的模式学习来获取对用户和物品的有效表示(Fan et al., 2023)，进而实现从离散的ID中捕捉用户对物品的偏好。然而，这种基于ID的推荐方法往往面临用户-物品交互数据稀疏、符号ID不含有语义信息等问题。

STAR(Sheng et al., 2021)、PeterRec(Yuan et al., 2022)、ALCDR(Zhao et al., 2023)等工作基于协同过滤思想，利用不同推荐平台之间共享的用户或物品信息，实现在不同领域间传输用户或物品级偏好的目的，这类方法可用于解决跨域推荐问题或冷启动推荐问题。但是，这些方法效果严重依赖不同领域间用户或物品的重叠，而在实际场景中，这种重叠假设几乎不成立(Yuan et al., 2023)，此外，用户和物品的ID仅作为一种索引符号，因而其本身无法为用户和物品的表示提供有价值的语义信息(Fan et al., 2023)。

对交互行为的研究表明，个体倾向于与其所属的群体相关联，群体的所有成员通常具有相似的购买行为并且彼此之间具有显著的影响力。例如，个体的职业背景会影响其购买行为(Naim, 2023)，人们倾向于购买与其职业和社会角色相符的产品和服务。这里，职业作为一种潜在驱动因素，使群体用户表现出特定的购买行为模式。同时，潜在的模式关联能够从多个角度捕捉用户偏好，进而提高模型的表现(Zhang et al., 2024)。因此，我们假设不同域用户-物品交互行为之间存在潜在的模式关联，如图1所示。即，存在一种将不同域的用户连接的潜在因素，在该因素驱动下，不同域间的部分用户产生出相近的用户-物品交互行为，我们把这些相近交互行为之间的公共特征称为交互行为语义模式。

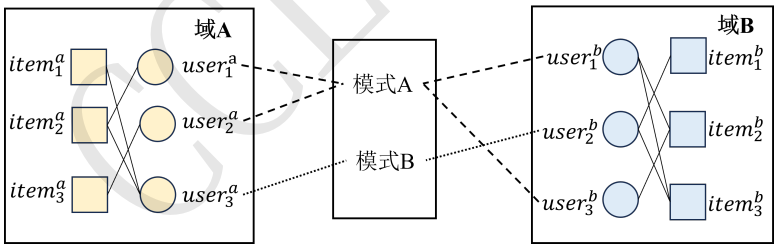



图 1. 交互行为语义模式表达潜在联系的示意

基于上述假设，本文提出了一种基于交互行为语义模式增强的ID推荐方法。该方法在目标域推荐任务中引入辅助域信息，基于图神经网络对辅助域和目标域信息进行联合编码表示，通过引入交互行为语义模式，将辅助域的用户-物品交互信息以及物品描述信息迁移至目标域，从而实现目标域ID推荐中的交互行为语义增强。本文的主要贡献如下：

（1）提出交互行为语义模式的假设，基于此构建基于交互行为语义模式的推荐方法，该方法能够在不依赖领域间用户或物品重叠的情况下，有效提升推荐效果。

（2）提出了基于交互行为语义模式增强的目标域的用户和物品表示模型PEIDR，通过图神经网络和预训练语言模型实现辅助域关联模式信息到目标域的迁移，从而增强目标域的用户和物品的语义表达。

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

（3）在8个公开数据集上进行了实验，结果表明，相比于当前的SOTA模型，本文提出的方法均获得了更好的表现，其Recall@20与NDCG@20分别具有3% ∼ 30%、1% ∼ 40%的提升，说明了本文提出假设的合理性及方法的有效性。

## 2 相关研究

基于单一领域数据实现的推荐系统长期面临数据稀疏和冷启动两大挑战，这促进了跨域推荐的出现和发展。跨域推荐的核心思想是利用从其他领域（辅助域）收集的信息来缓解目标领域中的数据稀疏和冷启动问题(Zang et al., 2022)。与传统推荐系统相比，跨领域推荐更为复杂。首先，考虑到两个域的用户和物品之间的关系，跨域推荐存在用用户重叠或非重叠等不同的推荐情况(Khan et al., 2017)。其次，跨领域推荐的推荐任务多种多样。例如，推荐用户和物品可能在同一领域，也可能在不同领域。最后，传统的推荐系统只需关注如何从历史交互记录中建立用户兴趣模型，对于跨领域推荐来说，除了在一个领域内对用户兴趣建模外，还需要考虑如何在不同领域之间传递知识。Man(2017)等人提出的EMCDR是一种经典的基于映射的方法，它通过学习映射函数在不同域之间转移重叠用户的偏好。Li(2020)等人提出的DDTCDR引入了一种潜在的正交映射来捕捉多个领域中的用户偏好，同时保留不同潜在空间中用户之间的关系。Xie(2022)等人提出的CCDR在不同领域之间对用户和物品属性进行对比学习，以实现多样化的知识转移。然而，这些仅基于ID关联记录的方法忽略了物品描述中的语义信息在推荐中的重要性。

SGL(Wu et al., 2021)将节点丢失(Node Dropout)、边缘丢失(Edge Dropout)和随机游走(Random Walk)应用于增强用户-物品二分图。它使用相同类型的增强算子生成两个增强图，并使用共享图LightGCN 编码器$f_\theta$(He et al., 2020)学习节点嵌入。Yu(2023)等人将对比学习引入推荐任务，利用用户-物品交互行为数据构建二分图，提出XSimGCL模型并证明了对比学习对图推荐模型的有益影响，XSimGCL模型通过优化对比损失InfoNCE获得比SGL(Wu et al., 2021)模型更为合理的用户与物品表征。本文的方法在XSimGCL模型的基础上再次优化用户与物品的特征表示，将辅助域中物品的语义信息融入目标域以获得带有语义信息的用户与物品的表示。

## 3 模型实现

### 3.1 问题定义

本文使用的符号及含义如表1所示。

| 符号名 | 意义 |
|---|---|
| $U_i^t$ | 目标域中用户$i$的特征表示 |
| $I_j^t$ | 目标域中物品$j$的特征表示 |
| $U^t$ | 目标域中用户特征表示的集合 |
| $I^t$ | 目标域中物品特征表示的集合 |
| $U_i'^t$ | 目标域中用户$i$融入语义的特征表示 |
| $I_j'^t$ | 目标域中物品$j$融入语义的特征表示 |
| $T_i$ | $item_i^a$的文本描述 |
| $S_i$ | 辅助域中物品文本描述$i$的特征表示 |
| $U_i^a$ | 辅助域中用户$i$的特征表示 |
| $I_j^a$ | 辅助域中物品$j$的特征表示 |
| $U^a$ | 辅助域中用户特征表示的集合 |
| $I^a$ | 辅助域中物品特征表示的集合 |
| $\overline{U}_j^a$ | 与$user_j^a$存在交互的物品的平均化表示 |
| $\overline{U}^a$ | $\overline{U}_j^a$的集合 |

表 1. 符号及其意义


(c) 2024 中国中文信息学会计算语言学专业委员会

在基于交互行为语义模式增强的ID推荐方法中，有来自于目标域和辅助域的两组用户集合，分别表示为$\mathrm{User}^{\mathrm{t}}=\{\mathrm{user}_1^{\mathrm{t}},\mathrm{user}_2^{\mathrm{t}},\mathrm{user}_3^{\mathrm{t}}...\mathrm{user}_{\mathrm{m}}^{\mathrm{t}}\}$和$\mathrm{User}^{\mathrm{a}}=\{\mathrm{user}_1^{\mathrm{a}},\mathrm{user}_2^{\mathrm{a}},\mathrm{user}_3^{\mathrm{a}}...\mathrm{user}_{\mathrm{n}}^{\mathrm{a}}\}$，其中m，n分别为目标域和辅助域中用户的个数；另有来自于目标域和辅助域的两组物品集合，分别表示为$\mathrm{Item}^{\mathrm{t}}=\{\mathrm{item}_1^{\mathrm{t}},\mathrm{item}_2^{\mathrm{t}},\mathrm{item}_3^{\mathrm{t}}\ldots\mathrm{item}_{\mathrm{p}}^{\mathrm{t}}\}$和$\mathrm{Item}^{\mathrm{a}}=\{\mathrm{item}_1^{\mathrm{a}},\mathrm{item}_2^{\mathrm{a}},\mathrm{item}_3^{\mathrm{a}}\ldots\mathrm{item}_{\mathrm{q}}^{\mathrm{a}}\}$，其中p，q分别为目标域和辅助域中物品的个数。$\mathrm{User}^{\mathrm{t}}$与$\mathrm{User}^{\mathrm{a}}$，$\mathrm{Item}^{\mathrm{t}}$与$\mathrm{Item}^{\mathrm{a}}$中均无重叠项，即，$\mathrm{User}^{\mathrm{a}}\cap\mathrm{User}^{\mathrm{t}}=\varnothing$；$\mathrm{Item}^{\mathrm{a}}\cap\mathrm{Item}^{\mathrm{t}}=\varnothing$。辅助域和目标域中的数据分别构成两个二分图，并以三元组$\langle\mathrm{User\text{-}ID},\mathrm{Item\text{-}ID},\mathrm{InteractionFlag}\rangle$的形式表示，其中，InteractionFlag={0,1}表示用户是否与物品存在交互。基于交互行为语义模式增强的ID推荐的目标是利用辅助域信息指导推荐模型在目标域上获得更为准确的物品推荐结果，该问题可形式化表示如下。

$$\max\sum_{i}^{m}\sum_{j}^{p}(2c_{ij}-1)x_{ij}. \tag{1}$$

其中，$c_{ij}=\{0,1\}$。$c_{ij}=0$表示$user_i^t$与$item_j^t$之间没有交互，$c_{ij}=1$表示$user_i^t$与$item_j^t$之间存在交互；$x_{ij}=\{0,1\}$。$x_{ij}=0$表示模型预测$user_i^t$与$item_j^t$之间没有交互，$x_{ij}=1$表示模型预测$user_i^t$与$item_j^t$之间存在交互。

### 3.2 模型概述

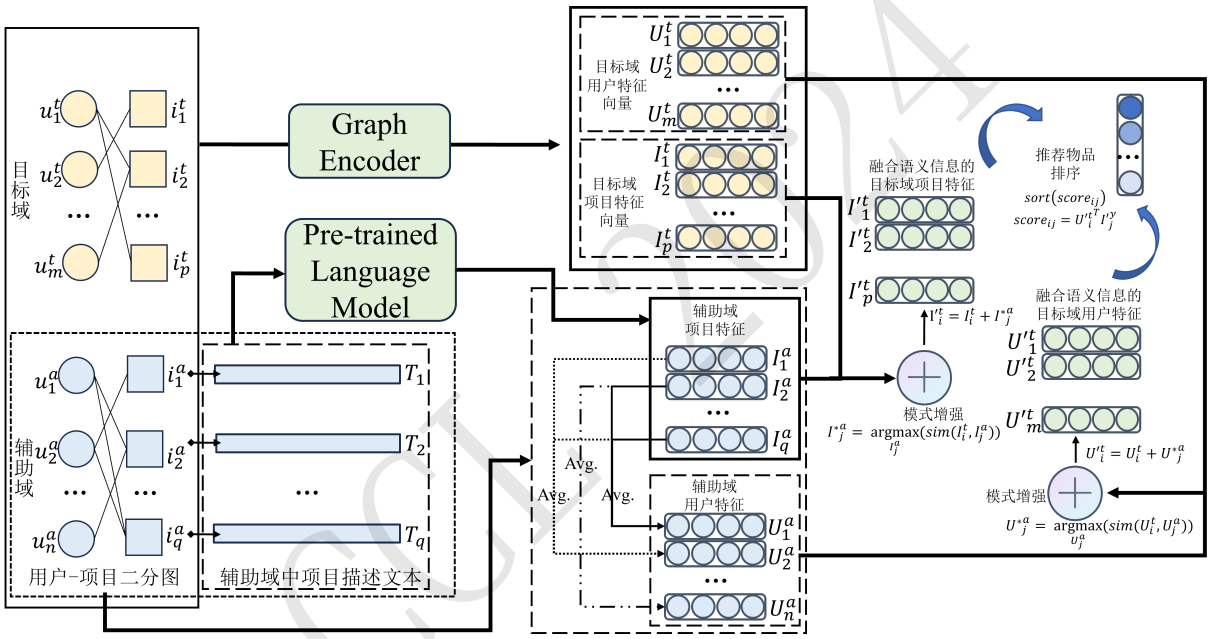



图 2. 模型结构图

本文设计了基于交互行为语义模式增强的目标域用户和物品表示模型PEIDR。为了引入语义信息，使用三元组$\langle\mathrm{Item\text{-}ID},\mathrm{Title},\mathrm{Describe}\rangle$间接建立辅助域中的物品描述信息(Describe)与用户的交互关系。PEIDR模型通过图神经网络和预训练语言模型实现辅助域关联模式信息到目标域的迁移，从而增强目标域的用户和物品的语义表达。另外，为了进一步提升推荐的质量和推荐的准确性，通过优化推荐损失的方式增强了用户和物品的特征表示。基于交互行为语义模式增强过程如算法1所示。

#### 3.2.1 交互行为语义模式增强

PEIDR模型的结构图如图2所示。通过图编码器$XSimGCL$(Yu et al., 2023)对目标域和辅助域的用户-物品二分图进行联合编码，得到目标域和辅助域中的用户与物品的嵌入向量。为了获取辅助域物品描述文本的语义信息，我们使用预训练语言模型BGE(Zhang et al., 2023)对辅助域中每个物品的文本描述$T_j\in T$，$j\in C_{\mathbb{N}}[1,q]$ 进行编码。

$$S_j=Encoder(T_j). \tag{2}$$


(c) 2024 中国中文信息学会计算语言学专业委员会

$S_j$被用来更新由图编码器产生的辅助域中用户与物品的特征。在辅助域中，为了获取符合用户偏好的语义特征，使用与该用户存在交互的所有物品特征平均化对其进行更新。其中，$\sum_j c_{ij}$表示$user_i^a$与辅助域中物品交互的数量，$i \in C_{\mathbb{N}}[1,n]$，$j \in C_{\mathbb{N}}[1,q]$。

$$U_i^a = \frac{\sum_j c_{ij} \cdot S_j}{\sum_j c_{ij}}, \quad c_{ij} = \{0,1\}. \tag{3}$$

目标域用户$user_i^t$更新后的表示向量$U_i'^t$由其更新前的表示向量$U_i^t$和与之距离最近的辅助域用户$user_j^a$的表示向量$U_j^{*a}$相加得到。获取融入语义信息的用户特征表示过程可形式化描述如下：

$$U_i'^t = U_i^t + U_j^{*a}. \tag{4}$$

$$U_j^{*a} = \underset{U_j^a}{\operatorname{argmax}}(sim(U_i^t, U_j^a)). \tag{5}$$

其中，$U_i^t \in U^t$, $U_j^a \in U^a$，$i \in C_{\mathbb{N}}[1,m]$，$j \in C_{\mathbb{N}}[1,n]$。

目标域物品$item_i^t$更新后的表示向量$I_i'^t$由其更新前的表示向量$I_i^t$和与之距离最近的辅助域物品$item_j^a$的表示向量$I_j^{*a}$相加得到。获取融入语义信息的物品特征表示过程可形式化描述如下：

$$I_i'^t = I_i^t + I_j^{*a}. \tag{6}$$

$$I_j^{*a} = \underset{I_j^a}{\operatorname{argmax}}(sim(I_i^t, I_j^a)). \tag{7}$$

$$I_i^a = Encoder(T_i). \tag{8}$$

其中，$I_i^t \in I^t$，$I_j^a \in I^a$，$i \in C_{\mathbb{N}}[1,p]$，$j \in C_{\mathbb{N}}[1,q]$。

基于用户与物品的特征向量点积计算排序结果为用户推荐TopK个物品，K=20。

$$score_{ij} = U_i'^{t\top} I_j'^t. \tag{9}$$

### 3.2.2 优化推荐损失

为了提高推荐系统整体性能，本文采用由超参数λ调节的推荐损失与对比损失联合的方式(Yu et al., 2023)进行模型优化。

$$L = L_{rec} + \lambda L_{cl}. \tag{10}$$

假设$\mathcal{B}$为当前批次，则$u$和$i$分别表示$\mathcal{B}$中的用户与物品序号。对比损失最小化使得图对比学习模型能够从ID的交互信息中学习到更均匀的物品表示。

$$L_{cl} = \sum_{u \in \mathcal{B}} -log\left(\frac{exp\left(z_u'^{\top} z_u^{l^*}/\tau\right)}{\sum_{i \in \mathcal{B}} exp\left(z_u'^{\top} z_i^{l^*}/\tau\right)}\right). \tag{11}$$

其中，$z_u'$和$z_u^{l^*}$是来自不同的层嵌入的用户特征，$l^*$表示与最终层对比的层序数，$\tau$是温度系数，控制对难负样本的惩罚强度。

考虑到将语义信息融入目标域中的用户和物品表示过程，推荐损失定义如下：

$$L_{rec} = -\sum_{(u,i) \in \mathcal{B}} log\Big(\sigma\left(e_{u_s}^{\top} e_{i_s} - e_{u_s}^{\top} e_{j_s}\right)\Big). \tag{12}$$

其中，$e_{u_s}$表示融合了语义信息的用户表示，$e_{i_s}$表示与用户$u$存在交互的且融合了语义信息的物品表示，$e_{j_s}$是随机采样的与用户不存在交互且融入语义信息的物品表示，$\sigma$是sigmoid函数。

CCL 2024

**算法 1** 基于交互行为语义模式增强算法

**Input:** The User-Item bipartite graphs from the target domain and auxiliary domains.
**Output:** Optimized Graph Encoder

$S \leftarrow Encoder(T)$
**for** each $U_j^a$ in $U^a$ **do**
  $//S_1, S_2, \ldots, S_k$为与$user_j^a$存在交互的物品表征，$k$为物品数量
  $\overline{U}_j^a \leftarrow mean(S_1, S_2, \ldots, S_k)$
  $\overline{U}^a.add(\overline{U}_j^a)$
**end for**
**for** $epoch$ in $range(max\_epoch)$ **do**
  $U^t$,$I^t \leftarrow model\_graph\_encoder$
  **for** each $U_i^t$ in $U^t$ **do**
    **for** each $TempU_j^a$ in $\overline{U}^a$ **do** $//TempU_j^a$为临时变量
      $U_j^{*a} \leftarrow \underset{U_j^a}{\text{argmax}}(sim(U_i^t, TempU_j^a))$
    **end for**
    $U_i'^t \leftarrow U_i^t + U_j^{*a}$
  **end for**
  **for** each $I_i^t$ in $I^t$ **do**
    **for** each $I_j^a$ in $I^a$ **do**
      //预训练语言模型对$item_j^a$的表示替换Graph Encoder对$item_j^a$的表示
      $I_j^a \leftarrow Encoder(T_j)$
      $I_j^{*a} \leftarrow \underset{I_j^a}{\text{argmax}}(sim(I_i^t, I_j^a))$
    **end for**
    $I_i'^t \leftarrow I_i^t + I_j^{*a}$
  **end for**
  $L \leftarrow L_{rec} + \lambda L_{cl}$
  $L.backward$
**end for**
**return** $U_i'^t$,$I_j'^t$
$score_{ij} \leftarrow U_i'^{t\top} I_j'^t$
$sort(score_{ij})$

## 4 实验

### 4.1 数据集

为了获得可靠且令人信服的结果，本文在8个公共数据集上进行了交叉实验评估。其中，Yelp2018[0](He et al., 2020)、Amazon-kindle[1](He and McAuley, 2016)、Alibaba-iFashion[2](Wu et al., 2021)、Amazon-Electronic[3](Ni et al., 2019)、Book Crossing Dataset[4](Hu et al., 2023)等是仅包含用户-物品ID交互信息的数据集，这里仅作为目标域数据集使用；MovieLens10M[5](Lee et al., 2022)、MovieLens20M[4](Yi et al., 2023)、MovieLens25M[4](Ji et al., 2023)和Citeulike-a[6]等是包含物品描述信息的数据集，这里逐一作为实验的辅助域数据集使用。目标域与辅助域中既没有重叠的用户也不存在重叠的物品。

[0] https://github.com/kuandeng/LightGCN/tree/master/Data/yelp2018
[1] https://github.com/Coder-Yu/SELFRec/tree/main/dataset/amazon-kindle
[2] https://github.com/Coder-Yu/SELFRec/tree/main/dataset/iFashion
[3] https://cseweb.ucsd.edu/~jmcauley/datasets/amazon/links.html
[4] https://grouplens.org/datasets/book-crossing/
[5] https://grouplens.org/datasets/movielens/
[6] https://github.com/js05212/citeulike-a

CCL 2024

| Dataset | User | Item | Interactions | Describe | Density |
|---|---|---|---|---|---|
| Yelp2018 | 31668 | 38048 | 1561406 | - | 1.3e-3 |
| Amazon-kindle | 138333 | 98572 | 1909965 | - | 1.4e-4 |
| Alibaba-iFashion | 300000 | 81614 | 1607813 | - | 7.0e-5 |
| Amazon-Electronics | 719376 | 159364 | 5460975 | - | 5.0e-6 |
| Book Crossing Dataset | 61844 | 142000 | 351646 | - | 4.0e-5 |
| MovieLens10M | 69878 | 10677 | 10000054 | ✓ | 1.3e-2 |
| MovieLens20M | 7120 | 14026 | 1048575 | ✓ | 1.0e-2 |
| MovieLens25M | 7045 | 22240 | 1048575 | ✓ | 7.0e-3 |
| Citeulike-a | 5551 | 16980 | 204986 | ✓ | 2.2e-3 |

表 2. 实验数据的详细信息

各数据集中的数据分别情况如表2所示。本文将各目标域数据集按7:1:2的比例分为三部分（训练集、验证集和测试集）进行独立验证实验，使用基于相关性的指标Recall@20 和基于排名的指标NDCG@20对推荐结果进行评价。

### 4.2 超参数设置

本研究中的实验采用自监督推荐的框架SELFRec[7]实现所有的模型和基线。选用预训练语言模型BGE(Zhang et al., 2023)作为语义提取器。在模型中设置向量维度为768维，并使用Adam优化器来最小化损失函数。学习率设置为$10^{-4}$，$L_2$正则化系数设置为$10^{-4}$，批次大小设置为2048，温度系数$\tau$设置为0.2，超参数 $\lambda$设置为0.2，$l^*$设置为1。在训练过程中，使用公式(10)损失函数来训练模型，并采用批量梯度下降算法进行优化。

### 4.3 对比基线

（1）LightGCN(He et al., 2020)：LightGCN 通过用户-物品交互图上线性传播用户和物品嵌入来学习用户和物品嵌入，并使用所有层学习到的嵌入的加权和作为最终嵌入。

（2）SGL(Wu et al., 2021)：自监督图学习范式。SGL-ND(Node Dropout)、SGL-ED(Edge Dropout)和SGL-RW(Random Walk)为图结构上三中不同的操作，根据不同的视图减少噪声的影响，达到图增强的目的。

（3）SGL-WA(Yu et al., 2022)：改进SGL变体，提出并非所有图增强都会产生积极影响，某些图增强（例如Node Dropout）可能会通过删除关键节点（例如中心）及其相关边来扭曲原始图，并证明在图对比学习中，对比损失的选择比图增强更有效果。

（3）SimGCL(Yu et al., 2022)：消除了无效的图形增强，而是通过向学习的表示添加均匀噪声来实现更有效表示。经验结果表明，这种基于噪声的增强可以获得更均匀的表示分布。

（4）XSimGCL(Yu et al., 2023)：优化SimGCL，是对比任务和推荐任务共享通道，简化计算过程。

### 4.4 实验结果

本文研究方法与基线方法进行了比较，实验结果汇总在表3中，每个指标的最佳分数用粗体表示，次优指标用下划线表示。实验结果表明，本文提出的方法在不同的数据集中都表现出了有效性。

### 4.5 结果分析

#### 4.5.1 交互行为语义模式的可转移性

通过分析8个数据集中涉及的推荐物品的领域，我们研究了潜在行为模式的领域可转移性。

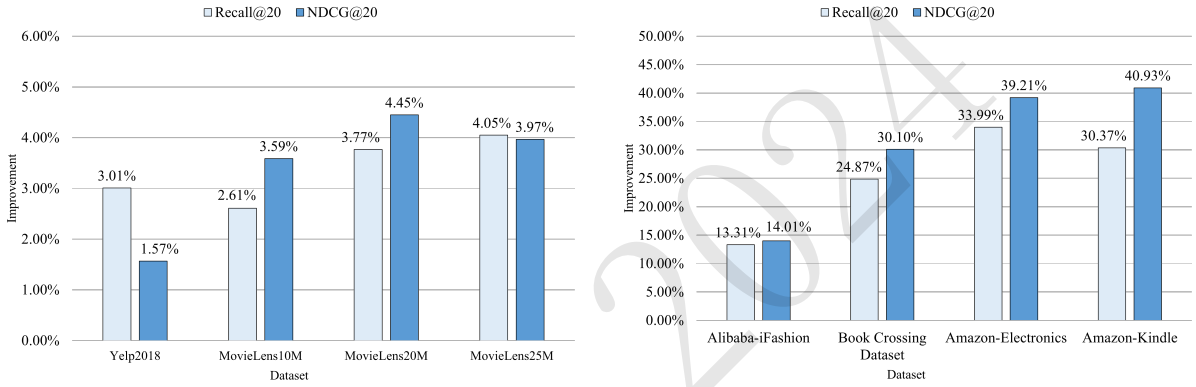
如图3所示，当我们使用Citeulike-a作为辅助域信息观察目标域中的推荐结果时，我们发现不同目标域之间的推荐效果存在差异。数据集Amazon-kindle、Amazon-Electronic和Book Crossing Dataset与数据集Yelp2018、Alibaba-iFashion、MovieLens10M、MovieLens20M和Movie-

[7] https://github.com/Coder-Yu/SELFRec


(c) 2024 中国中文信息学会计算语言学专业委员会

| Dataset | Evaluation | Method | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | LightGCN | SGL-ND | SGL-ED | SGL-RW | SGL-WA | SimGCL | XSimGCL | PEIDR(Ours) |
| Yelp2018 | Recall@20 | 0.0622 | 0.0658 | 0.0668 | 0.0644 | 0.0653 | 0.0719 | 0.0722 | **0.0744** |
| | NDCG@20 | 0.0504 | 0.0538 | 0.0549 | 0.0530 | 0.0544 | 0.0601 | 0.0604 | **0.0613** |
| Amazon-Kindle | Recall@20 | 0.2033 | 0.2020 | 0.2084 | 0.2088 | 0.2068 | 0.2071 | 0.2114 | **0.2756** |
| | NDCG@20 | 0.1284 | 0.1307 | 0.1341 | 0.1345 | 0.1330 | 0.1341 | 0.1382 | **0.1948** |
| Alibaba-iFashion | Recall@20 | 0.1053 | 0.0993 | 0.1062 | 0.1053 | 0.1028 | 0.1119 | 0.1143 | **0.1295** |
| | NDCG@20 | 0.0505 | 0.0484 | 0.0514 | 0.0512 | 0.0501 | 0.0548 | 0.0559 | **0.0637** |
| Amazon-Electronics | Recall@20 | 0.0545 | 0.0665 | 0.0688 | 0.0692 | 0.0681 | 0.0698 | 0.0704 | **0.0943** |
| | NDCG@20 | 0.0352 | 0.0465 | 0.0496 | 0.0497 | 0.0489 | 0.0493 | 0.0521 | **0.0725** |
| MovieLens10M | Recall@20 | 0.2747 | 0.2642 | 0.2735 | 0.2726 | 0.2741 | 0.2854 | 0.3002 | **0.3080** |
| | NDCG@20 | 0.3345 | 0.3208 | 0.3302 | 0.3295 | 0.3309 | 0.3373 | 0.3591 | **0.3720** |
| MovieLens20M | Recall@20 | 0.2305 | 0.2368 | 0.2446 | 0.2438 | 0.2456 | 0.2381 | 0.2615 | **0.2713** |
| | NDCG@20 | 0.2938 | 0.3032 | 0.3116 | 0.3135 | 0.3119 | 0.2909 | 0.3262 | **0.3407** |
| MovieLens25M | Recall@20 | 0.2256 | 0.2360 | 0.2423 | 0.2425 | 0.2419 | 0.2398 | 0.2606 | **0.2711** |
| | NDCG@20 | 0.2894 | 0.2995 | 0.3094 | 0.3076 | 0.3076 | 0.2937 | 0.3257 | **0.3386** |
| Book Crossing Dataset | Recall@20 | 0.0426 | 0.0454 | 0.0481 | 0.0483 | 0.0452 | 0.0431 | 0.0444 | **0.0603** |
| | NDCG@20 | 0.0235 | 0.0263 | 0.0274 | 0.0278 | 0.0261 | 0.0242 | 0.0249 | **0.0361** |

表 3. PEIDR和基线在8个数据集上的比较结果




(a) (b)

图 3. PEIDR与最新SOTA模型的提升对比

Lens25M相比，前者在推荐效果上的提升更加明显，这说明辅助域中的推荐项属性与目标域中的推荐项属性越接近，辅助域信息的传递增强效果越好。

### 4.5.2 交互行为语义模式的有效性

为了证明潜在行为模式增强的有效性，进行了有无行为语义模式增强的消融实验，消融实验的超参数设置与对比试验中的PEIDR方法相同。如表4所示的实验结果表明，引入行为语义模式增强能够有效地提升模型的推荐结果。

| 数据集 | Recall@20 | | | NDCG@20 | | |
|---|---|---|---|---|---|---|
| | PEIDR(Ours) | 无行为语义模式增强 | Improvement | PEIDR(Ours) | 无行为语义模式增强 | Improvement |
| Yelp2018 | 0.0744 | 0.0729 | 2.04% | 0.0613 | 0.0602 | -1.88% |
| Amazon-Kindle | 0.2756 | 0.2330 | 18.28% | 0.1948 | 0.1500 | -29.87% |
| Alibaba-iFashion | 0.1295 | 0.1172 | 10.50% | 0.0637 | 0.0568 | -12.16% |
| Amazon-Electronics | 0.0943 | 0.0840 | 12.34% | 0.0725 | 0.0635 | -14.24% |
| MovieLens10M | 0.3080 | 0.2914 | 5.72% | 0.3720 | 0.3557 | -4.59% |
| MovieLens20M | 0.2713 | 0.2631 | 3.13% | 0.3407 | 0.3359 | -1.44% |
| MovieLens25M | 0.2711 | 0.2623 | 3.38% | 0.3386 | 0.3333 | -1.57% |
| Book Crossing Dataset | 0.0603 | 0.0515 | 16.94% | 0.0361 | 0.0319 | -13.12% |

表 4. 消融实验结果


(c) 2024 中国中文信息学会计算语言学专业委员会

### 4.5.3 交互行为语义模式的鲁棒性

为了验证交互行为语义模式的鲁棒性，我们分别使用带文本描述信息的MovieLens10M、MovieLens20M和MovieLens25M作为辅助域数据集进行交叉验证。为了避免样本数量差异对实验结果的影响，我们随机选取与Citeulike-a相同数量的用户－物品交互记录作为辅助域数据。实验结果如表5至表7所示。每个指标的最佳分数用粗体表示，次优指标用下划线表示，表中最后一列表示该数据集的最佳结果相比次优结果的提升率。实验结果进一步表明，使用不同的辅助域数据在本文PEIDR模型框架下均产生了不同程度推荐效果的提升，表明本文提出的交互行为语义模式具有鲁棒性。

| Dataset | Evaluation | Method | | | | | | | | Improvement |
|---|---|---|---|---|---|---|---|---|---|---|
| | | LightGCN | SGL-ND | SGL-ED | SGL-RW | SGL-WA | SimGCL | XSimGCL | PEIDR(Ours) | |
| Yelp2018 | Recall@20 | 0.0622 | 0.0658 | 0.0668 | 0.0644 | 0.0653 | 0.0719 | 0.0722 | **0.0742** | 2.80% |
| | NDCG@20 | 0.0504 | 0.0538 | 0.0549 | 0.0530 | 0.0544 | 0.0601 | 0.0604 | **0.0613** | 1.49% |
| Amazon-Kindle | Recall@20 | 0.2033 | 0.2020 | 0.2084 | 0.2088 | 0.2068 | 0.2071 | 0.2114 | **0.2740** | 29.62% |
| | NDCG@20 | 0.1284 | 0.1307 | 0.1341 | 0.1345 | 0.1330 | 0.1341 | 0.1382 | **0.1938** | 40.23% |
| Alibaba-iFashion | Recall@20 | 0.1053 | 0.0993 | 0.1062 | 0.1053 | 0.1028 | 0.1119 | 0.1143 | **0.1284** | 12.30% |
| | NDCG@20 | 0.0505 | 0.0484 | 0.0514 | 0.0512 | 0.0501 | 0.0548 | 0.0559 | **0.0631** | 12.81% |
| Amazon-Electronics | Recall@20 | 0.0545 | 0.0665 | 0.0688 | 0.0692 | 0.0681 | 0.0698 | 0.0704 | **0.0935** | 32.84% |
| | NDCG@20 | 0.0352 | 0.0465 | 0.0496 | 0.0497 | 0.0489 | 0.0493 | 0.0521 | **0.0718** | 37.85% |
| citeulike-a | Recall@20 | 0.2361 | 0.2511 | 0.2620 | 0.2628 | 0.2591 | 0.2413 | 0.2590 | **0.2806** | 6.78% |
| | NDCG@20 | 0.1929 | 0.2064 | 0.2148 | 0.2158 | 0.2128 | 0.2045 | 0.2150 | **0.2334** | 8.13% |
| Book Crossing Dataset | Recall@20 | 0.0426 | 0.0454 | 0.0481 | 0.0483 | 0.0452 | 0.0431 | 0.0444 | **0.0546** | 13.12% |
| | NDCG@20 | 0.0235 | 0.0263 | 0.0274 | 0.0278 | 0.0261 | 0.0242 | 0.0249 | **0.0325** | 17.10% |

表 5. 使用数据集MovieLens10M作为辅助域的增强结果

| Dataset | Evaluation | Method | | | | | | | | Improvement |
|---|---|---|---|---|---|---|---|---|---|---|
| | | LightGCN | SGL-ND | SGL-ED | SGL-RW | SGL-WA | SimGCL | XSimGCL | PEIDR(Ours) | |
| Yelp2018 | Recall@20 | 0.0622 | 0.0658 | 0.0668 | 0.0644 | 0.0653 | 0.0719 | 0.0722 | **0.0742** | 2.77% |
| | NDCG@20 | 0.0504 | 0.0538 | 0.0549 | 0.0530 | 0.0544 | 0.0601 | 0.0604 | **0.0614** | 1.59% |
| Amazon-Kindle | Recall@20 | 0.2033 | 0.2020 | 0.2084 | 0.2088 | 0.2068 | 0.2071 | 0.2114 | **0.2752** | 30.19% |
| | NDCG@20 | 0.1284 | 0.1307 | 0.1341 | 0.1345 | 0.1330 | 0.1341 | 0.1382 | **0.1951** | 41.14% |
| Alibaba-iFashion | Recall@20 | 0.1053 | 0.0993 | 0.1062 | 0.1053 | 0.1028 | 0.1119 | 0.1143 | **0.1284** | 12.33% |
| | NDCG@20 | 0.0505 | 0.0484 | 0.0514 | 0.0512 | 0.0501 | 0.0548 | 0.0559 | **0.0631** | 12.84% |
| Amazon-Electronics | Recall@20 | 0.0545 | 0.0665 | 0.0688 | 0.0692 | 0.0681 | 0.0698 | 0.0704 | **0.0936** | 32.88% |
| | NDCG@20 | 0.0352 | 0.0465 | 0.0496 | 0.0497 | 0.0489 | 0.0493 | 0.0521 | **0.0719** | 37.97% |
| citeulike-a | Recall@20 | 0.2361 | 0.2511 | 0.2620 | 0.2628 | 0.2591 | 0.2413 | 0.2590 | **0.2822** | 7.39% |
| | NDCG@20 | 0.1929 | 0.2064 | 0.2148 | 0.2158 | 0.2128 | 0.2045 | 0.2150 | **0.2343** | 8.57% |
| Book Crossing Dataset | Recall@20 | 0.0426 | 0.0454 | 0.0481 | 0.0483 | 0.0452 | 0.0431 | 0.0444 | **0.0604** | 25.16% |
| | NDCG@20 | 0.0235 | 0.0263 | 0.0274 | 0.0278 | 0.0261 | 0.0242 | 0.0249 | **0.0350** | 25.89% |

表 6. 使用数据集MovieLens20M作为辅助域的增强结果

| Dataset | Evaluation | Method | | | | | | | | Improvement |
|---|---|---|---|---|---|---|---|---|---|---|
| | | LightGCN | SGL-ND | SGL-ED | SGL-RW | SGL-WA | SimGCL | XSimGCL | PEIDR(Ours) | |
| Yelp2018 | Recall@20 | 0.0622 | 0.0658 | 0.0668 | 0.0644 | 0.0653 | 0.0719 | 0.0722 | **0.0742** | 2.80% |
| | NDCG@20 | 0.0504 | 0.0538 | 0.0549 | 0.0530 | 0.0544 | 0.0601 | 0.0604 | **0.0612** | 1.32% |
| Amazon-Kindle | Recall@20 | 0.2033 | 0.2020 | 0.2084 | 0.2088 | 0.2068 | 0.2071 | 0.2114 | **0.2757** | 30.41% |
| | NDCG@20 | 0.1284 | 0.1307 | 0.1341 | 0.1345 | 0.1330 | 0.1341 | 0.1382 | **0.1952** | 41.24% |
| Alibaba-iFashion | Recall@20 | 0.1053 | 0.0993 | 0.1062 | 0.1053 | 0.1028 | 0.1119 | 0.1143 | **0.1284** | 12.31% |
| | NDCG@20 | 0.0505 | 0.0484 | 0.0514 | 0.0512 | 0.0501 | 0.0548 | 0.0559 | **0.0632** | 12.97% |
| Amazon-Electronics | Recall@20 | 0.0545 | 0.0665 | 0.0688 | 0.0692 | 0.0681 | 0.0698 | 0.0704 | **0.0934** | 32.67% |
| | NDCG@20 | 0.0352 | 0.0465 | 0.0496 | 0.0497 | 0.0489 | 0.0493 | 0.0521 | **0.0718** | 37.87% |
| citeulike-a | Recall@20 | 0.2361 | 0.2511 | 0.2620 | 0.2628 | 0.2591 | 0.2413 | 0.2590 | **0.2861** | 8.85% |
| | NDCG@20 | 0.1929 | 0.2064 | 0.2148 | 0.2158 | 0.2128 | 0.2045 | 0.2150 | **0.2391** | 10.81% |
| Book Crossing Dataset | Recall@20 | 0.0426 | 0.0454 | 0.0481 | 0.0483 | 0.0452 | 0.0431 | 0.0444 | **0.0577** | 19.64% |
| | NDCG@20 | 0.0235 | 0.0263 | 0.0274 | 0.0278 | 0.0261 | 0.0242 | 0.0249 | **0.0343** | 23.62% |

表 7. 使用数据集MovieLens25M作为辅助域的增强结果

CCL 2024

## 5 结论

本文针对基于ID的推荐方法经常面临用户-物品交互数据稀疏、符号ID缺失语义信息等问题，提出了不同领域用户-物品交互行为之间存在潜在的模式关联的假设，并基于此提出了一种基于交互行为语义模式增强的ID推荐方法。实验结果表明，相比当前SOTA模型，该方法能够在无跨域用户和物品重叠的情况下实现语义信息迁移，并表现出更好的推荐效果，说明了本文提出假设的合理性及方法的有效性；在多个领域ID推荐任务上的实验表明，相近领域之间的交互行为模式具有更好的可迁移性，辅助域与目标域越接近，引入辅助域信息对目标域ID推荐结果提升越明显。

## 致谢

本文受到国家自然科学基金(U1908216)的资助。

## 参考文献

Wenqi Fan, Zihuai Zhao, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Jiliang Tang, and Qing Li. 2023. Recommender systems in the era of large language models (llms). arxiv 2023. *arXiv preprint arXiv:2307.02046*.

Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*, pages 507–517.

Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. 2020. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 639–648.

Zheng Hu, Shi-Min Cai, Jun Wang, and Tao Zhou. 2023. Collaborative recommendation model based on multi-modal multi-view attention network: Movie and literature cases. *Applied Soft Computing*, page 110518.

Chao Huang, Huance Xu, Yong Xu, Peng Dai, Lianghao Xia, Mengyin Lu, Liefeng Bo, Hao Xing, Xiaoping Lai, and Yanfang Ye. 2021. Knowledge-aware coupled graph neural network for social recommendation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 4115–4122.

Yitong Ji, Aixin Sun, Jie Zhang, and Chenliang Li. 2023. A critical study on data leakage in recommender system offline evaluation. *ACM Transactions on Information Systems*, 41(3):1–27.

Muhammad Murad Khan, Roliana Ibrahim, and Imran Ghani. 2017. Cross domain recommender systems: a systematic literature review. *ACM Computing Surveys (CSUR)*, 50(3):1–34.

Jae-woong Lee, Seongmin Park, Joonseok Lee, and Jongwuk Lee. 2022. Bilateral self-unbiased learning from biased implicit feedback. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 29–39.

Pan Li and Alexander Tuzhilin. 2020. Ddtcdr: Deep dual transfer cross domain recommendation. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 331–339.

Tong Man, Huawei Shen, Xiaolong Jin, and Xueqi Cheng. 2017. Cross-domain recommendation: An embedding and mapping approach. In *IJCAI*, volume 17, pages 2464–2470.

Arshi Naim. 2023. Consumer behavior in marketing patterns, types, segmentation. *European Journal of Economics, Finance and Business Development*, 1(1):1–18.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 188–197.


(c) 2024 中国中文信息学会计算语言学专业委员会

Xiang-Rong Sheng, Liqin Zhao, Guorui Zhou, Xinyao Ding, Binding Dai, Qiang Luo, Siran Yang, Jingshan Lv, Chi Zhang, Hongbo Deng, et al. 2021. One model to serve all: Star topology adaptive recommender for multi-domain ctr prediction. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 4104–4113.

Shoujin Wang, Longbing Cao, Yan Wang, Quan Z Sheng, Mehmet A Orgun, and Defu Lian. 2021. A survey on session-based recommender systems. *ACM Computing Surveys (CSUR)*, 54(7):1–38.

Jiancan Wu, Xiang Wang, Fuli Feng, Xiangnan He, Liang Chen, Jianxun Lian, and Xing Xie. 2021. Self-supervised graph learning for recommendation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 726–735.

Ruobing Xie, Qi Liu, Liangdong Wang, Shukai Liu, Bo Zhang, and Leyu Lin. 2022. Contrastive cross-domain recommendation in matching. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4226–4236.

Jing Yi, Xubin Ren, and Zhenzhong Chen. 2023. Multi-auxiliary augmented collaborative variational auto-encoder for tag recommendation. *ACM Transactions on Information Systems*, 41(4):1–25.

Junliang Yu, Hongzhi Yin, Xin Xia, Tong Chen, Lizhen Cui, and Quoc Viet Hung Nguyen. 2022. Are graph augmentations necessary? simple graph contrastive learning for recommendation. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 1294–1303.

Junliang Yu, Xin Xia, Tong Chen, Lizhen Cui, Nguyen Quoc Viet Hung, and Hongzhi Yin. 2023. Xsimgcl: Towards extremely simple graph contrastive learning for recommendation. *IEEE Transactions on Knowledge and Data Engineering*.

Guanghu Yuan, Fajie Yuan, Yudong Li, Beibei Kong, Shujie Li, Lei Chen, Min Yang, Chenyun Yu, Bo Hu, Zang Li, et al. 2022. Tenrec: A large-scale multipurpose benchmark dataset for recommender systems. *Advances in Neural Information Processing Systems*, 35:11480–11493.

Zheng Yuan, Fajie Yuan, Yu Song, Youhua Li, Junchen Fu, Fei Yang, Yunzhu Pan, and Yongxin Ni. 2023. Where to go next for recommender systems? id- vs. modality-based recommender models revisited. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 2639–2649. Association for Computing Machinery.

Tianzi Zang, Yanmin Zhu, Haobing Liu, Ruohan Zhang, and Jiadi Yu. 2022. A survey on cross-domain recommendation: taxonomies, methods, and future directions. *ACM Transactions on Information Systems*, 41(2):1–39.

Peitian Zhang, Shitao Xiao, Zheng Liu, Zhicheng Dou, and Jian-Yun Nie. 2023. Retrieve anything to augment large language models. *arXiv preprint arXiv:2310.07554*.

Xiaokun Zhang, Bo Xu, Zhaochun Ren, Xiaochen Wang, Hongfei Lin, and Fenglong Ma. 2024. Disentangling id and modality effects for session-based recommendation. *arXiv preprint arXiv:2404.12969*.

Yi Zhao, Chaozhuo Li, Jiquan Peng, Xiaohan Fang, Feiran Huang, Senzhang Wang, Xing Xie, and Jibing Gong. 2023. Beyond the overlapping users: cross-domain recommendation via adaptive anchor link learning. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*, pages 1488–1497.

CCL 2024

# 基于双层语义映射的大语言模型辅助古汉语事件抽取半自动标注框架

**卫聪聪, 李炜, 冯振冰, 邵艳秋***
北京语言大学，信息科学学院，
国家语言资源监测与研究平面媒体中心，
北京市海淀区学院路15号，100083
{weicongcong0214, zbfengblcu, yqshao163}@163.com, liweitj47@blcu.edu.cn

## 摘要

尽管自然语言处理技术（NLP）在现代语言事件抽取任务（EE）上已有较为成熟的解决方案，但针对古汉语事件抽取的研究却受限于标注数据匮乏和文本语义复杂等挑战。因而我们提出使用当前取得巨大成功的大语言模型（LLMs）来辅助人类标注员进行数据标注。为了应对LLMs在古汉语上存在的训练不足、语义理解能力欠缺的问题，我们提出了一种基于双层语义映射的LLMs辅助古汉语事件抽取半自动标注框架，利用古汉语的现代汉语译文，结合事件语义学理论及语义依存分析技术，为LLMs提供丰富的语义信息表示，从而进一步将语义依存关系逐步映射为具体的事件信息。经过人类标注员的审核反馈，有效克服了现有NLP工具和LLMs在古汉语事件抽取标注时的局限。实验结果表明，我们的方法不仅提高了古汉语事件抽取标注的准确性和效率，而且减少了对专业人员的依赖和人工标注工作量，为低资源语言标注实践提供了新的方法论，探索了大模型时代数据标注的新方向。

**关键词：** 事件抽取；大语言模型；语义依存分析；事件语义学

# A Semi-automatic Annotation Framework for Event Extraction in Classical Chinese Assisted by Large Language Models Based on Dual-Layer Semantic Mapping

**Congcong Wei, Wei Li, Zhenbing Feng, Yanqiu Shao***
Information Science School, Beijing Language and Culture University,
Language Resources Monitoring and Research Center,
15 Xueyuan Road, HaiDian District, Beijing, 100083
{weicongcong0214, zbfengblcu, yqshao163}@163.com, liweitj47@blcu.edu.cn

## Abstract

While Natural Language Processing(NLP) techniques have advanced for modern languages, Classical Chinese Event Extraction (EE) research is limited by data scarcity and semantic complexity. Thus, we propose using successful Large Language Models (LLMs) to aid human annotators. To address LLMs' training shortcomings and semantic comprehension in Classical Chinese, we introduce a semi-automatic annotation framework assisted by LLMs Based on Dual-Layer Semantic Mapping. This framework uses Modern Chinese translations of Classical Chinese texts, combined with event semantics theory and semantic dependency analysis, to enhance the semantic representations for LLMs. It methodically converts these semantic dependencies into detailed event information. The process is iteratively refined by human annotators, thus effectively addressing the limitations of existing NLP tools and LLMs in annotating Classical Chinese events. Experimental evidence shows that our method improves

* 通讯作者 Corresponding Author

第二十三届中国计算语言学大会论文集，第588页-第599页，太原，中国，2024年7月25日至28日。
卷1：主会论文
(c) 2024 中国中文信息学会计算语言学专业委员会

Classical Chinese EE annotation's accuracy and efficiency, lowers reliance on experts, and introduces innovative annotation methods for low-resource languages, forging new avenues in data annotation during the LLM era.

**Keywords:** Event Extraction , Large Language Models , Semantic Dependency Analysis , Event Semantics

## 1 引言

事件抽取（EE）是自然语言处理（NLP）的一个重要任务，旨在从大规模非结构化文本中自动识别事件及其相关实体和属性(Chang et al., 2000)。事件抽取通常包括以下几个子任务：事件检测、事件关系分析和事件论元抽取等（如图1所示）。**事件检测**旨在确定文本中所描述的事件属于哪一种或多种事件类型，包括识别文本中的触发词，并判断这些词所属的事件类型(Wang et al., 2020)；**事件关系分析**则识别和理解文本中不同事件之间的关系，包括逻辑及时序关系等(Wang et al., 2022)；**事件论元抽取**是从文本中识别和抽取与特定事件相关的实体及角色(Ma et al., 2022)。

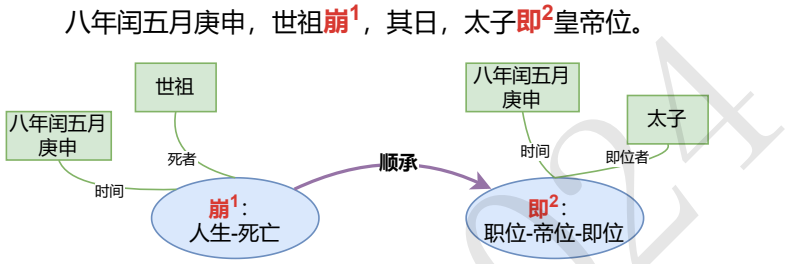



图 1: 事件抽取任务示意图。识别“崩”与“即”为触发词，对应“死亡”与“即位”事件类型，事件关系为顺承。论元抽取识别出时间（八月闰五月庚申）和主体（世祖、太子）。

针对古汉语的事件抽取是一个深具意义的研究领域，不仅有助于构建古汉语事件的本体知识，为深入挖掘和理解古汉语事件信息提供了理论框架，而且对推动事件知识图谱的构建(Guan et al., 2021)、智能问答系统的开发(Meng et al., 2021)、大语言模型（LLMs）知识评测(Yu et al., 2023)等多个领域起着重要作用。此外，研究古籍中的事件对于保护和传承历史文化遗产也至关重要，有助于更深刻地理解和传播历史文化。

然而，当前取得较大成功的事件抽取工具均依赖高质量、数量大的数据集资源。尽管NLP技术在现代语言处理上已有较成熟的解决方案(Li et al., 2021)，但由于古汉语的显著差异，其事件抽取任务无法直接利用现代汉语的数据资源，当前古汉语事件抽取标注面临着诸多挑战，数据集资源尤为匮乏。

具体来说，现有的NLP工具多数基于现代汉语训练，对于古汉语的特殊语法和语义理解不足，尤其缺乏专门针对古汉语的开源语义分析工具。同时，由于古汉语的歧义性、模糊性以及独特的语法现象（如省略和词类活用），从古汉语文本中准确提取事件信息变得极为困难。自从以GPT-3.5、GPT-4等为代表的模型发布以来，NLP领域进入了一个大模型时代(Zhao et al., 2023)。因而有学者也开始尝试直接使用这些模型处理古汉语问题(Wang et al., 2023)或事件抽取任务(Gao et al., 2023)，但由于古汉语训练数据的缺乏，LLMs在理解古汉语的语义上存在局限，因而在事件抽取任务上也无法达到理想效果。

尽管直接使用LLMs处理古汉语存在局限，我们提出可以通过为其提供结构更加清晰的现代汉语文本和丰富的语义信息来提高LLMs在古汉语领域的语义解析能力。因而我们提出了一种基于双层语义映射的LLMs辅助半自动标注框架（如图2所示）。具体来说，双层语义映射是指，（1）**第一层语义映射**：将古汉语翻译为现代汉语，利用**语义依存分析**技术将现代汉语映射为丰富的语义信息表示，旨在解决古汉语语义复杂性的问题。（2）**第二层语义映射**：在第一层映射的基础上，结合**事件语义学理论**及**语义依存关系框架**建立事件语义解析框架并设计提示

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

词，进一步利用LLMs对这些已分析的文本进行事件抽取标注，将语义依存关系逐步映射为具体事件信息。最后，**半自动标注框架**则结合了LLMs的自动标注和人类标注员的审核反馈，既提高了效率，又保证了标注的准确性和可靠性。

为验证方法的有效性，我们进行了交叉验证标注实验。结果表明，我们的方法不仅提高了古汉语事件抽取标注的准确性和效率，而且减少了对专业人员的依赖和标注人员的工作量，为古汉语事件抽取标注提供了一种新的解决方案。总的来说，我们的贡献如下：

- 我们提出引入LLMs用于古汉语文本标注，并开发半自动标注框架，结合了LLMs的自动标注能力和人类标注员的审核反馈，显著提升了低资源语言标注的准确率与效率，探索了大模型时代的数据标注的新方向。
- 我们通过双层语义映射方法、结合语义依存分析技术和事件语义学理论，为LLMs提供了丰富的语义信息表示，有效缓解了其在古汉语上存在的训练不足、语义理解能力欠缺的问题，从而更深入地分析古汉语文本中的事件信息。
- 我们从标注准确率、标注时间等多个维度验证了所提方法的可行性与有效性，展现了我们方法在应对古汉语文本标注挑战方面的优势，为低资源语言的标注实践提供了新的理论框架，推动了古汉语文本处理的发展。

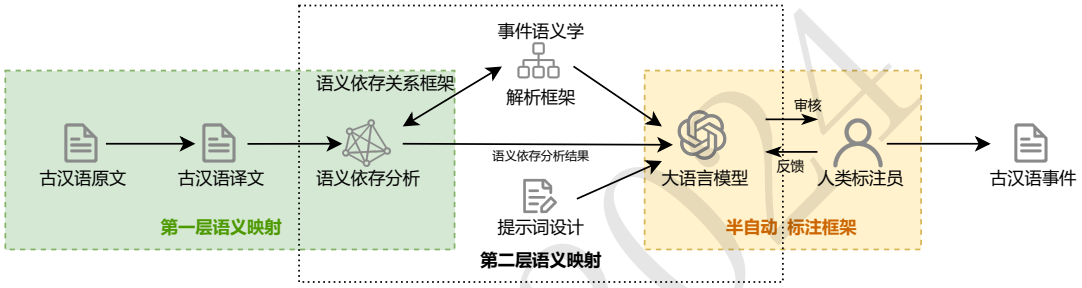



图 2: 基于双层语义映射的LLMs辅助古汉语事件抽取半自动标注框架示意图

## 2 相关工作

### 2.1 古汉语事件抽取

近年来，部分学者已经开始对古汉语事件抽取进行了初步的研究。这些研究大致可以分为基于模式匹配、机器学习和神经网络的方法。例如，李等(李章超et al., 2020)的研究为例，他们使用Fillmore的框架理论来研究《左传》中的战争事件，并通过模式匹配和条件随机场模型对相关事件进行识别和抽取。另外，刘等(刘忠宝et al., 2020)则在《史记》语料上，采用了BERT模型和LSTM-CRF模型进行历史事件和其组成元素的抽取。喻等(喻雪寒et al., 2021)学者研究了《左传》中的战争句，并利用RoBERTa-CRF模型进行事件抽取。更进一步，他们(喻雪寒et al., 2023)在基于神经网络的方法中融入了机器阅读理解模式，为事件抽取提供了新的思路，并在《左传》和《史记》上进行了实验。总之，这些方法在特定文本上有效且改进了古汉语事件抽取，但由于它们依赖于现代汉语优化模型，在处理古汉语的复杂性、特殊语法和语义时存在局限。

### 2.2 大模型事件抽取

目前，已有学者尝试探索LLMs用于事件抽取的可行性。Wei等(Wei et al., 2023)的研究从零样本信息抽取（IE）的角度出发，探讨了ChatGPT在未经特定训练情况下的信息提取效果。他们将零样本IE任务转化为一个包含两阶段框架的多轮问答问题（ChatIE）。在三个IE任务（实体-关系三元组提取、命名实体识别、事件抽取）的广泛评估中，ChatIE在六个数据集上的表现令人印象深刻。而Gao等(Gao et al., 2023)的工作探索了在长尾和复杂场景（即包含多个事件的文本）下ChatGPT用于事件抽取的效果，研究发现，ChatGPT在处理长尾和复杂场景（例如多事件文本）时的性能只有特定任务模型的一半。总之，虽然LLMs在事件抽取方面展现出潜力，但在处理特定类型的语言结构时仍存在显著挑战。


(c) 2024 中国中文信息学会计算语言学专业委员会

### 2.3 事件语义学研究

事件语义学是语义学的一个重要理论分支，核心是将事件作为理解句子意义的基本元素，强调事件在句子中的角色。Davidson(Davidson, 1967)首次在行动句（action sentences）中引入“事件”论元，认为动词不仅关系到主语和宾语，还关联到一个表示实际动作的“事件”。随后，事件语义学的研究领域不断扩展。如研究者们尝试区分不同类型的事件（如状态、过程和结果），并探讨它们是如何通过语言结构被表达和理解的(Maienborn, 2011)。事件的因果关系，即事件之间的相互作用和影响，也是研究的重点之一。Talmy(Talmy, 2000)对此进行了深入探讨，提出了初始，持续等因果关系概念，强调了事件语义在理解语言中的动态过程的重要性。随着计算语言学的发展，事件语义学在NLP和语言资源构建中也受到关注。例如，有学者利用事件语义学分析日语和汉语中的存在句，建立它们的事件结构模型，以此来处理语言信息(熊苇渡, 2020)。此外，一些学者还利用事件语义学理论，探索如何计算复杂句子中事件的语义距离，进而帮助构建语言知识词典(马腾and 詹卫东, 2014)。总之，事件语义学的发展不仅加深了我们对语言语义的理解，而且对事件抽取的框架提供了理论基础。

## 3 方法

我们使用图3来简要阐述我们的方法流程。图中主要分为四大部分，输入为二十四史语料，首先进行数据准备和预处理工作，其次构建事件语义解析框架，对古汉语译文进行语义依存分析（第一层语义映射），并设计提示词，再使用LLMs标注分析（第二层语义映射），然后人类进行审核映射（半自动标注框架），最终输出为二十四史事件抽取语料。

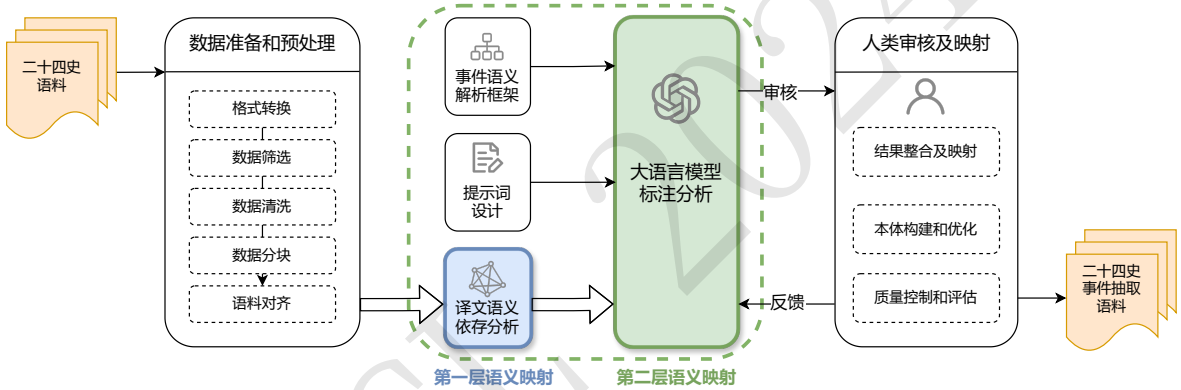



图 3: 方法流程图。展示从数据预处理到二十四史事件抽取语料的生成过程，其中蓝色为第一层语义映射，将译文映射为丰富语义信息；绿色（含虚线部分）为第二层语义映射，将语义信息进一步映射为事件信息。

### 3.1 数据准备和预处理

我们选择《二十四史》作为主要研究对象，这套权威文献详细记录了从黄帝时期（约公元前2550年）至明朝崇祯十七年（1644年）的中国历史，展现了不同朝代、各色人物的历史纪实。其中的“本纪”和“列传”章节，作为各史书的核心内容，是我们研究的重点。为确保语料的完整性和质量，我们采用了文白对照的《二十四史》精华版（共12册）[0]作为语料来源，并将对其进行预处理。如图4所示为数据预处理示例。具体来说：

- **格式转换**：将数据源格式转换为便于标注的格式。互联网公开的电子书籍多为equb等格式，我们需要将其转换为txt或json文本格式，以便进行标注。同时，为了更好地检索和理解文档，我们会在数据中加入文档ID和篇章信息。

- **数据筛选**：筛选出我们的目标语料。主要是以人物或篇章为检索单位，筛选出原文和对应译文数据。

[0]来源网址: http://www.happydot.top/9151.html


(c) 2024 中国中文信息学会计算语言学专业委员会

- **数据清洗**：删去无关数据内容，如文本中的多余符号等。这一步骤使用正则化表达式进行规则匹配，清洗数据中的一些噪音，以降低错误率。
- **数据分块**：将大批量文本数据切分为更小批次，方便处理，存储及排查错误。对于过长的文本，我们会在保持上下文连贯性的前提下进行分块，以提高效率。
- **语料对齐**：将古汉语原文与现代汉语译文进行句子级别的对齐，为语义映射打下基础。由于译文通常比原文长，这一步骤有助于确保分析结果能准确地对应到原文。


```
{
      "doc_id": "1史记卷八十一列传第二十一廉颇蔺相如列传",
      "sentence_id": "2",
      "total_sentences": "235",
      "original_text": "赵惠文王十六年，廉颇为赵将伐齐，大破之，取阳晋，拜为上卿，以勇气闻于诸侯。",
      "translation": "赵惠文王十六年，廉颇率领赵军攻打齐国，大败齐军，攻占阳晋，以军功官拜上卿，他也就以勇敢无畏而闻名于诸侯各国。"
}
```


图 4: 数据预处理后的古汉语语料JSON格式示例。展示文档编号（doc_id）、句子编号（sentence_id）、总句数（total_sentence）、原文（original_text）及对应翻译（translation）。

### 3.2 事件语义解析框架构建

在探索事件抽取和语义依存分析的联系部分，我们结合了事件语义学理论和语义依存关系框架[1]，旨在将语义依存分析标签映射到事件语义标签上，以构建一个全面事件语义解析框架。

**事件语义学**认为“事件”是一种具有确定时空属性并能够整合参与者功能的特定实体；强调事件的时空特性和可感知性，即事件作为世界中的实体，在空间和时间中有其特定的位置，并且在实现方式上可能有所不同（多样化）。例如，句子“七月丙寅，始皇死于沙丘平台。”中，此事件发生在特定时间（七月丙寅）和地点（沙丘平台），其中始皇为参与者，其死亡为事件核心。此事件是历史上真实发生的，可以被感知。进一步，死亡方式可能有多种，如自然死亡、疾病或他杀等，这表明同一事件（死亡）可以不同方式发生（如图5b所示）。

而**语义依存分析**旨在理解和分析语句中词语之间的语义关系。语义依存分析规范认为事件是以一个核心谓词为中枢论元，一个或多个相关的论元为周边的语义组合所反映的客观现实。例如，上例句子的语义依存分析中，核心谓词“死”连接多个相关论元。其中，始皇是当事，意味着他是事件的主体；“七月丙寅”代表时间，“沙丘平台”代表地点，共同构建了这一历史事件的背景（如图5a所示）。

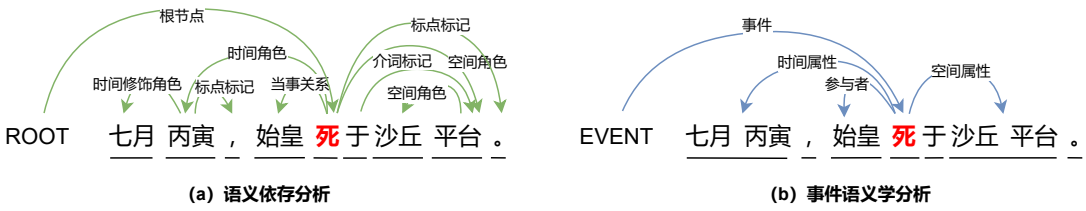



图 5: 语义依存分析（a）与事件语义学分析（b）对比分析

通过比较，可以发现两者都视事件为具有时空属性的客观现实，包含核心元素和相关参与者，且均具有可感知性和多样性。我们可以认为语义依存关系的部分框架和事件语义解析框架

[1]https://csdp-doc.readthedocs.io/zh-cn/latest/


(c) 2024 中国中文信息学会计算语言学专业委员会

相关，且存在映射关系（如图6所示）。事件抽取着眼于文本中的实体行为，涉及特定词语触发的事件结构。这一结构包括：触发词、参与者、情境、情态及事件关系。具体来看：

- **触发词（trigger）**：事件发生的核心词，常为动词或名词，对应语义依存分析的核心谓词。
- **参与者（participant）**：事件参与的主体和客体，对应语义依存分析中的主体角色和客体角色。
- **情境（situation）**：事件发生的时间、地点等背景要素，对应语义依存分析的情境角色。
- **情态（modal）**：人对事件的主观态度，如是否真实发生，对应语义依存分析中的否定和情态标记。
- **事件关系（event_relations）**：事件间的联系，对应语义依存分析的事件关系。

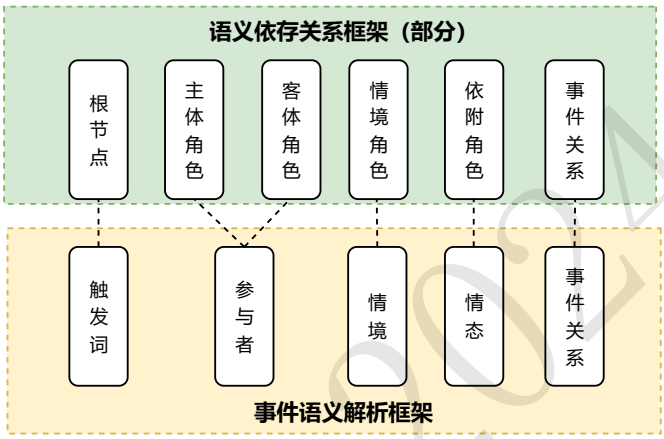



图 6: 语义依存关系框架（部分）与事件语义映射关系示意图。展示语义依存关系框架中各角色与事件语义解析框架中相应元素的对应关系。

### 3.3 第一层语义映射：译文语义依存分析

由于古汉语语义的复杂性，目前缺乏专门针对古汉语文本的语义依存分析工具，且现代NLP工具主要基于现代语言语料训练，对古汉语的特殊语法和语义处理能力有限。相反，现代汉语清晰的语法结构有助于NLP工具更准确地进行语义依存分析，从而提高事件信息提取的准确性，具有可行性。

因而我们使用语义依存分析工具[2]对古汉语译文进行语义依存分析，作为双层语义映射方法的第一层映射。语义依存分析是理解和抽取事件信息的关键，对译文进行语义依存分析对于事件抽取不仅是必要的，而且是提供深度语义理解的关键步骤。它能够有效地辅助抽取古汉语文本中的复杂事件关系，准确识别触发词和解析事件结构，为构建精确的事件本体和理解事件动态提供坚实基础。具体来说:

- **确定事件触发词**：事件的触发词通常是动词或名词，它们表明了发生了什么动作或事件。语义依存分析能够识别句子中的核心动词和名词，而这些词往往是事件的触发词。如句子“八年闰五月庚申日，孝武帝驾崩，当天，太子业登皇帝位。”中，“驾崩”和“登”就是两个事件的触发词。
- **解析事件论元**：语义依存分析中，每个词语被赋予特定的语义角色，而这些语义角色中就已包含了时间、空间、主体、客体等事件发生的要素。如在上述例句中，“孝武帝”和“太子业”是两个事件的主体。

[2]https://hanlp.hankcs.com/


(c) 2024 中国中文信息学会计算语言学专业委员会

- **揭示事件间的关系**：语义依存分析不仅识别词语之间的表层句法关系，还深入探究它们之间的语义关系，通过分析语义角色和关系，能够揭示事件之间的逻辑关系，如因果、顺承或并列等。如在上句中，“太子业登皇帝位”紧接着“孝武帝驾崩”发生，形成了一个顺承关系。

### 3.4 第二层语义映射：大语言模型标注分析

LLMs的通用性在于其能将多种任务转化为文本生成。通过设计简洁的提示词或提供少量示例，它能执行众多NLP任务，无需为特定任务构建专门的数据集或模型。利用这一特性，我们提出使用LLMs进行数据初步标注。尽管LLMs在直接处理古汉语方面存在局限，但我们通过提供结构更加清晰的现代汉语文本和丰富的语义信息，既能发挥LLMs高效的语义解析能力，又能规避其在处理古文时的不足，提高标注效率。

因而，在第一层语义映射及事件语义解析框架的基础上，我们使用LLMs对已进行语义依存分析的现代汉语文本进行事件抽取标注，将语义依存关系映射为具体事件信息（如图8所示）。这一过程包括两个阶段：提示词设计和模型标注。

- **提示词设计**：根据“角色+ 任务（详情）+ 目标+ 格式/内容要求+ 示例”的模板创建提示词（如图7所示）。“角色”界定了LLMs的角色，“任务”说明了具体的任务背景和细节，“目标”是模型应输出的结果，而“格式/内容”要求则规定了输出的具体形式。这个过程是迭代的，通过反复测试和评估模型输出，不断调整提示词以获取稳定的结果。

- **思维链及少样本学习**：为了提高LLMs分析的准确性，我们还采用了思维链（CoT）(Wei et al., 2022)技术和少样本学习方法(Brown et al., 2020)。利用思维链技术，提示词模拟了人类处理事件语义信息的逻辑顺序，这一过程逐步分析和提取语义依存关系框架中的关键成分。首先识别触发词以确定事件的核心，接着将其主体及客体角色对应到事件参与者上，然后通过提取时间和地点信息，确定事件的具体情境，处理情态词以捕捉事件的主观态度，并构建事件之间的逻辑和因果关系。整个映射过程确保了事件信息的准确性和完整性。最后，少样本学习则通过展示具体的标签映射和输出格式示例，使模型迅速适应并执行新任务。

- **模型标注**：我们利用gpt-4-1106-preview模型作为基础，通过API进行自动化标注。这包括将译文的语义依存分析结果输入模型，并将这些关系映射到事件分析上，即把语义依存关系转换为具体的事件信息。

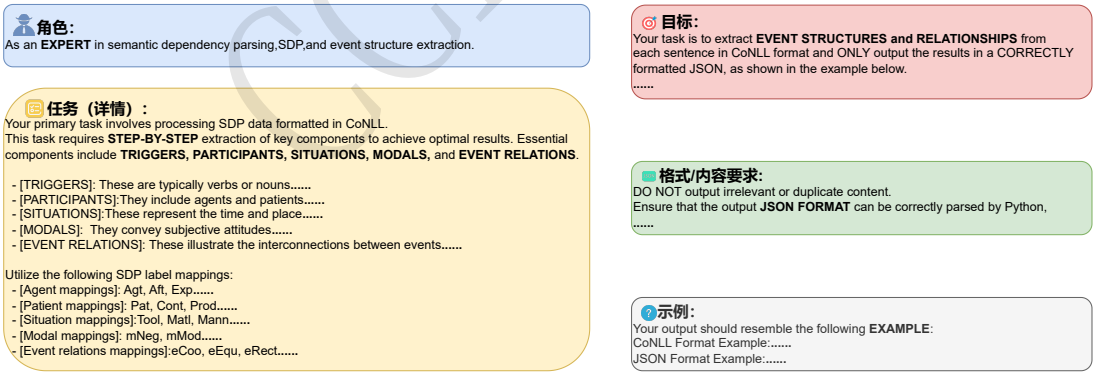



图 7: 输入给LLMs的提示词示意图。展示提示词的各部分，“角色+ 任务（详情）+ 目标+ 格式/内容要求+ 示例”，以及思维链、少样本学习方法。

### 3.5 人类审核及映射

尽管LLMs在许多方面表现出色，但在处理复杂和特定的语言任务时仍可能产生错误。而人类标注员可以识别并纠正这些错误，更进一步提高标注结果的准确性。因而在半自动标注框架

```
{
    "translation": "三年，蒙骜攻打韩国，夺取了十三个城邑。",
    "events": [
    {"event_id": "E1","trigger": "攻打","participant": [{"Agent": {"Agt": "蒙骜"}},
{"Patient": {"Datv": "韩国"}}],"situation": [{"Time": "三年"}],"modal": []},
    {"event_id": "E2","trigger": "夺取","participant": [{"Agent": {"Agt": "蒙骜"}},
{"Patient": {"Pat": "城邑"}}],"situation": [{"Time": "了"},{"Quan": "十三个"}],"modal": []}
    ],
    "event_relations": [{"source_event": "E1","target_event": "E2","relation_type":
"eResu"}]
}
```


图 8: LLMs输出的JSON格式展示。呈现古汉语事件的结构化信息，含译文（translation）、事件触发词（trigger）、参与者角色（participant）、情境（situation）及事件关系（event_relation）

中，LLMs首先进行初步标注，随后人类标注员对结果进行审核反馈。通过对照古今对齐语料，标注结果可准确映射回古汉语原文，有效弥补了模型在处理古汉语特有问题时的不足。

此外，基于LLMs的初步标注结果，人类标注员可以进一步进行**本体构建和优化**。结合语义依存分析提供的深层语义知识和LLMs的语义理解能力，我们能有效抽取古汉语文本中复杂事件关系，为构建精确事件本体和理解事件动态提供坚实基础，帮助人类标注员发现一些新的事件类型和事件关系，从而进一步优化本体知识（如图9所示）。

如在句子“契丹寇澶州，（安重霸）以临阵忸怩，为景延广所诛。”中，传统观点未将“忸怩”这类情感词视为事件类型。然而，通过分析结果，我们发现“忸怩”对事件发展有重要作用，呈现契丹军队攻击澶州与安重霸临阵畏缩间的转折关系，以及导致其被景延广处决的因果关系。因此，我们的方法能挖掘古汉语中特定事件类型，尤其是影响事件发展的人物情感词汇。




图 9: 我们方法在发现新事件类型和关系中的应用示例。图中句子分析揭示了传统未视为事件类型的情感词“忸怩”在事件发展中重要性，显示了古汉语文本的特色。

## 4 实验

### 4.1 设置

本实验旨在验证我们提出的方法对古汉语事件抽取标注任务的准确率、标注效率及标注难度的影响。实验设计包括**两个对照组**，一组在原文上直接进行标注（原文标注组），另一组则应用我们方法进行标注（辅助标注组）。具体实验任务涉及对古汉语原文句子进行触发词识别、事件论元和事件关系标注，标注难度依次递增，标签选取基于语义依存关系。

为了全面评估我们方法的效果，我们采取了**多维度**的测量方法。评价指标包括准确率（Precision）、召回率（Recall）和F1值（F1 Score），同时考量标注时间。首先，我们计算对照组在三个任务上的准确率、召回率和F1值，以全面评估标注员标注效果。这些指标计算的基准（baseline）是来自多位专家的综合标注结果。其次，我们统计了对照组的平均标注时间。具体操作中，我们在限定时间内观察对照组的标注句子数量，以此来计算平均标注时间。各指标的具体概念和计算方法如下：

- **准确率(Precision)**：指正确标注的正例数量与所有标注为正例的数量之比。准确率度量了标注人员在标注触发词、事件论元和事件关系时的准确性。

$$\text{准确率（Precision）} = \frac{\text{正确标注的正例数目}}{\text{所有标注为正例的数目}} \tag{1}$$

- **召回率(Recall)**：指正确标注的正例数量与所有应该标注为正例的数量之比。召回率度量了标注人员从文本中正确标注事件元素的能力，召回率高意味着标注人员漏标的情况较少。

$$\text{召回率（Recall）} = \frac{\text{正确标注的正例数目}}{\text{所有应该标注为正例的数目}} \tag{2}$$

- **F1值(F1 Score)**：指准确率和召回率的调和平均数。F1值被用来综合评价标注人员在识别和标注文本中事件元素（如触发词、事件论元和事件关系）的效果。

$$\text{F1值（F1 Score）} = 2 \times \frac{\text{准确率} \times \text{召回率}}{\text{准确率} + \text{召回率}} \tag{3}$$

- **平均标注时间**：是标注一定数量句子所需时间的平均值，用于衡量标注效率和难度。这反映了标注人员在给定时间内完成标注任务的速度。

$$\text{平均标注时间（分钟/句）} = \frac{\text{标注所用总时间（分钟）}}{\text{标注的句子数量}} \tag{4}$$

同时，为了保证实验的一致性和合理性，我们从二十四史的63篇文档中随机筛选了320句古汉语原文作为标注实验语料。这些语料在句子总长度和标签数量上进行了严格控制，以确保句子难度的均匀分布。**标注人员的选择**也是实验设计的关键环节。我们选取了具备一定古汉语基础知识的人员，并将标注分为**试标注**（20句）和**正式标注**（300句）两个阶段（如图11所示）。试标注环节的目的在于确定标注人员的知识水平，以确保参与实验的标注人员的水平相当。正式标注环节则分为**两轮**进行，每轮各150句，同一标注人员需采用不同对照组方式对句子进行多次**交叉标注**。此外，我们制定了标注规范，并对标注人员进行了初步培训，从而减少了人为因素对实验结果的潜在偏差，增强了实验结果的可靠性。

### 4.2 结果及分析

在本次实验中，我们对提出的基于双层语义映射的LLMs辅助古汉语事件抽取半自动标注框架进行了全面的实验评估。在表1和图10中我们给出了实验结果，可以看到我们的方法在提高古汉语事件抽取标注的准确性、完整性和效率方面都具有**显著优势**。特别是在事件关系标注这一复杂任务上，提升效果尤为突出，显示了其在处理古汉语文本中复杂语义信息时的高效性和准确性。

具体而言，在触发词识别任务中，我们的方法相较于直接标注方法，尤其在召回率方面表现尤其优异，这表明我们的方法能够更全面地识别文本中的关键信息，减少了遗漏重要触发词的情况。而对于事件论元及事件关系标注任务，我们的方法同样显示出了更高的准确率和召回率，尤其是事件关系标注的提升最为显著，F1值几乎提升了一倍，这证明了我们方法在理解和分析复杂事件关系方面的有效性。

同时，图10显示，**我们的方法显著减少了平均标注时间，提高了标注效率**。这一点在两位标注员的数据上都得到了验证，其中标注员B的效率提升百分比高于标注员A，这可能与个体差异有关，但整体趋势一致表明我们的方法能有效缩短标注所需时间，提升标注效率。此外，平均标注时间也从侧面反映出**标注任务的难度**，一般情况下我们认为标注时间少则意味着标注任务较为简单。

此外，**不同任务本身的复杂度导致了各评价指标之间的显著差异**。具体来说，触发词识别任务的评价指标相对较高，因为它主要涉及到在文本中识别显而易见的动词或名词，较为简单。相反，事件论元识别和事件关系标注任务的评价指标较低，这一差异归因于这两项任务的复杂性。事件论元识别任务要求识别与特定事件相关的多种元素，这些元素可能包括古汉语中特有的人名、地名和官职等实体。这要求标注人员拥有对相关文化背景的深入了解。而对于事件关系标注任务则要求标注人员不仅理解文本中明确表达的信息，还要对古汉语中隐含的复杂语义关系有深刻的理解。因此，该任务的评价指标较低反映了其在认知和分析层面的高难度。


(c) 2024 中国中文信息学会计算语言学专业委员会

表 1: 对照组在不同标注任务上的效果对比

| 对照组 | 任务 | 标注员A | | | 标注员B | | |
|---|---|---|---|---|---|---|---|
| | | 准确率 | 召回率 | F1值 | 准确率 | 召回率 | F1值 |
| 原文标注组 | 触发词识别 | 78.95 | 77.74 | 78.34 | 65.48 | 81.18 | 72.49 |
| | 事件论元标注 | 46.42 | 38.80 | 42.27 | 58.22 | 53.11 | 55.55 |
| | 事件关系标注 | 26.35 | 25.36 | 25.85 | 26.28 | 28.72 | 27.45 |
| 辅助标注组 | 触发词识别 | 77.08 | 85.61 | **81.12** | 76.74 | 80.49 | **78.57** |
| | 事件论元标注 | 58.92 | 57.05 | **57.97** | 58.93 | 55.85 | **57.35** |
| | 事件关系标注 | 44.90 | 46.81 | **45.83** | 37.61 | 38.86 | **38.23** |

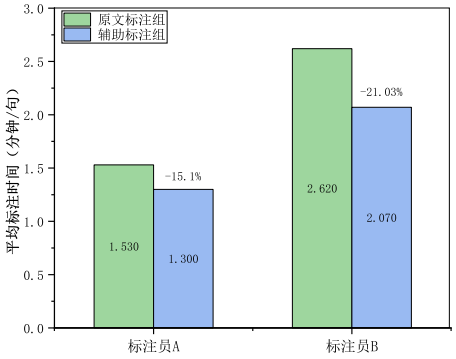



图 10: 对照组在标注时间上的对比

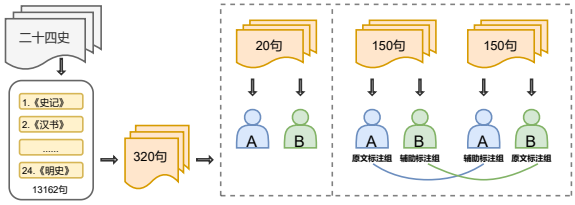



图 11: 实验标注阶段示意图。如图标注分为试标注（20句）和正式标注（300句）两个阶段，其中正式标注分为两轮（150句）进行，标注人员AB需采用不同对照组方式对句子进行交叉标注。

## 5 讨论

### 5.1 优点

我们的方法引入了一种基于双层语义映射的LLMs辅助半自动标注框架，专门针对古汉语事件抽取的数据标注。这一方法的核心在于融合了NLP工具和LLMs的优势，并结合人类标注员的审核反馈，有效缓解了古汉语文本的语义复杂性问题。具体来说：

（1）**通过提供丰富语义帮助LLMs进行分析，提高准确性**：古汉语文本的复杂性一直是NLP的一大挑战。我们的方法不仅保留了古文的丰富语义，还通过精确和可靠的分析，降低了文本的复杂度。同时，借助现代汉语译文的语义依存分析结果，我们的方法能够减少歧义，减少误标的可能性。

（2）**借助自动化工具和LLMs，提升标注效率与质量**：通过利用自动化工具和LLMs快速生成初步标注数据，结合人类专家的审核和反馈，我们的方法显著提高了标注的效率和质量。这种双管齐下的策略确保了数据标注的准确性和可靠性。

（3）**减少专业背景依赖和标注工作量**：传统的NLP工具在处理古汉语文本时面临诸多挑战，尤其是语义层面的复杂性和专业知识的需求。我们的方法通过双层语义映射，不仅减少了对背景知识的限制，还减少了工作量，使得标注过程更加容易和普遍适用。

### 5.2 局限

虽然我们的方法在提升古汉语事件抽取准确性等方面取得了效果，但仍存在一些局限。首先，**译文的选择直接影响语义依存分析的准确性**。依赖质量不高的译文进行事件抽取标注，可能导致标注结果质量不高。为此，我们强化数据预处理操作，选用出版书籍语料作为高质量译文。同时，我们对古汉语原文和现代汉语译文进行了对齐。

**语义依存分析局限于句子内，缺乏上下文信息**。由于我们分析的是古汉语历史文本篇章，而语义依存分析通常局限于单个句子内部，可能无法完全捕捉跨句子或跨篇章的事件关系。面对这一局限，我们在数据预处理阶段中加入上下文索引信息（如文档id和篇章位置），每一个句子都能明确索引到具体的文档中的上下文信息，为标注人员提供了参考。

**LLMs的标注结果的质量受限于提示词设计**。如果提示词设计不当，模型可能无法生成准确的标注结果。此外，由于LLMs在处理复杂输入时可能出现“幻觉（hallucination）(Li et al.,


(c) 2024 中国中文信息学会计算语言学专业委员会

2023)”现象，即产生与实际情况不符或逻辑混乱的回答。针对这一局限，我们在提示词设计阶段使用数据进行反复测试，观察评估LLMs的输出结果，不断调试提示词以便获得稳定的预期输出结果。同时使用了一些提示技术（如思维链(Wei et al., 2022)和少样本学习(Brown et al., 2020)）。

总而言之，对于我们方法的局限性，一个解决方法是结合**半自动标注框架**，其中LLMs提供初步标注，而人类标注员负责审核反馈，这种方法结合了LLMs的效率和人类标注员的深入理解，可以最大化事件抽取的准确性和可靠性。

## 6 结论与未来工作

我们通过使用基于双层语义映射的LLMs辅助半自动标注框架，有效地解决了古汉语事件抽取标注中的诸多挑战。我们首先将古汉语文本翻译为现代汉语，结合事件语义学理论和语义依存分析技术将译文映射为丰富的语义信息来解决古汉语语义复杂性的问题，进而利用LLMs进行深入分析，从而提高了对低资源语言标注的准确性。此外，我们结合了LLMs的自动标注能力与人类标注员的审核反馈，提高了标注效率及结果的可靠性。我们的方法在解决古汉语文本的复杂性等挑战方面表现出色，为低资源语言处理提供了新的方法论，探索了LLM时代数据标注的新方向。未来，我们将继续从数据质量和提示词设计方面优化我们的方法，并将其应用于更广泛的应用场景如历史文献的数字化分析等，以进一步推动数字人文学科和NLP领域的发展。

## 致谢

本成果受国家自然科学基金项目（62306045，61872402），北京语言大学校级项目（中央高校基本科研业务费专项资金）（18ZDJ03），北京语言大学梧桐创新平台项目（21PT04）基金资助。

## 参考文献

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Heng-Hsou Chang, Yau-Hwang Kuo, and Jang-Pong Hsu. 2000. An event-driven and ontology-based approach for the delivery and information extraction of e-mails. In *2000 International Symposium on Multimedia Software Engineering, ISMSE 2000, Taipei, Taiwan, December 11-13, 2000*, pages 103–109. IEEE Computer Society.

Donald Davidson. 1967. The logical form of action sentences. In Nicholas Rescher, editor, *The Logic of Decision and Action*, pages 81–95. University of Pittsburgh Press.

Jun Gao, Huan Zhao, Changlong Yu, and Ruifeng Xu. 2023. Exploring the feasibility of chatgpt for event extraction. *CoRR*, abs/2303.03836.

Saiping Guan, Xueqi Cheng, Long Bai, Fujun Zhang, Zixuan Li, Yutao Zeng, Xiaolong Jin, and Jiafeng Guo. 2021. What is event knowledge graph: A survey. *CoRR*, abs/2112.15280.

Qian Li, Jianxin Li, Jiawei Sheng, Shiyao Cui, Jia Wu, Yiming Hei, Hao Peng, Shu Guo, Lihong Wang, Amin Beheshti, and Philip S. Yu. 2021. A compact survey on event extraction: Approaches and applications.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 6449–6464. Association for Computational Linguistics.


(c) 2024 中国中文信息学会计算语言学专业委员会

Yubo Ma, Zehao Wang, Yixin Cao, Mukai Li, Meiqi Chen, Kun Wang, and Jing Shao. 2022. Prompt for extraction? PAIE: prompting argument interaction for event argument extraction. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 6759–6774. Association for Computational Linguistics.

Claudia Maienborn. 2011. 34. event semantics.

Fanqi Meng, Wenhui Wang, and Jingdong Wang. 2021. Visual analysis of the research status of intelligent question answering system. In *2021 13th International Conference on Computational Intelligence and Communication Networks (CICN)*, pages 145–149.

Leonard Talmy. 2000. Toward a cognitive semantics, vol. 1: Concept structuring systems.

Xiaozhi Wang, Ziqi Wang, Xu Han, Wangyi Jiang, Rong Han, Zhiyuan Liu, Juanzi Li, Peng Li, Yankai Lin, and Jie Zhou. 2020. MAVEN: A massive general domain event detection dataset. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 1652–1671. Association for Computational Linguistics.

Xiaozhi Wang, Yulin Chen, Ning Ding, Hao Peng, Zimu Wang, Yankai Lin, Xu Han, Lei Hou, Juanzi Li, Zhiyuan Liu, Peng Li, and Jie Zhou. 2022. MAVEN-ERE: A unified large-scale dataset for event coreference, temporal, causal, and subevent relation extraction. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 926–941. Association for Computational Linguistics.

Zhaoji Wang, Shirui Zhang, Xuetao Zhang, and Renfen Hu. 2023. 古汉语通假字资源库的构建及应用研究(the construction and application of an Ancient Chinese language resource on tongjiazi). In Maosong Sun, Bing Qin, Xipeng Qiu, Jing Jiang, and Xianpei Han, editors, *Proceedings of the 22nd Chinese National Conference on Computational Linguistics*, pages 535–546, Harbin, China, August. Chinese Information Processing Society of China.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, Yong Jiang, and Wenjuan Han. 2023. Zero-shot information extraction via chatting with chatgpt. *CoRR*, abs/2302.10205.

Jifan Yu, Xiaozhi Wang, Shangqing Tu, Shulin Cao, Daniel Zhang-li, Xin Lv, Hao Peng, Zijun Yao, Xiaohan Zhang, Hanming Li, Chunyang Li, Zheyuan Zhang, Yushi Bai, Yantao Liu, Amy Xin, Nianyi Lin, Kaifeng Yun, Linlu Gong, Jianhui Chen, Zhili Wu, Yunjia Qi, Weikai Li, Yong Guan, Kaisheng Zeng, Ji Qi, Hailong Jin, Jinxin Liu, Yu Gu, Yuan Yao, Ning Ding, Lei Hou, Zhiyuan Liu, Bin Xu, Jie Tang, and Juanzi Li. 2023. Kola: Carefully benchmarking world knowledge of large language models. *CoRR*, abs/2306.09296.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Z. Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jianyun Nie, and Ji rong Wen. 2023. A survey of large language models. *ArXiv*, abs/2303.18223.

刘忠宝, 党建飞, and 张志剑. 2020. 《史记》历史事件自动抽取与事理图谱构建研究. 图书情报工作.

喻雪寒, 何琳, and 徐健. 2021. 基于roberta-crf的古文历史事件抽取方法研究. 数据分析与知识发现, 5(7):26–35, 7.

喻雪寒, 何琳, and 王献琪. 2023. 基于机器阅读理解的古文事件抽取研究. 情报学报, 42(3):316–326, 3.

李章超, 李忠凯, and 何琳. 2020. 《左传》战争事件抽取技术研究. 图书情报工作, 64(7):20–29, 4.

熊苇渡. 2020. 面向语言信息处理的存现句事件语义学类型研究. 面向语言信息处理的存现句事件语义学类型研究.

马腾and 詹卫东. 2014. 基于事件语义距离的v1-v2述结式判别研究2014年2月27日. 计算机工程与应用.


(c) 2024 中国中文信息学会计算语言学专业委员会

# 基于文本风格迁移的中文性别歧视文本去毒研究

**彭健　左家莉*　谭景璇　万剑怡　王明文**
江西师范大学　计算机信息工程学院　江西　南昌　330022
Email: {pengjian2022, zjl, tanjingxuan, mwwang}@jxnu.edu.cn, wanjianyi@aliyun.com

**摘要**

网络社交媒体平台存在一定程度的性别歧视言论，阻碍了互联网健康和社会文明发展。文本风格迁移技术可以减轻文本中的性别歧视，在英语等语言上已有不少研究。但在中文领域，由于缺乏数据集而导致相关研究较少。此外，由于中文语义信息丰富、语言表达多样而导致性别歧视言论毒性的表现形式多样，现有的方法多采用单一文本风格迁移模型因而效果不佳。因此，本文提出了一个基于文本风格迁移的中文性别歧视文本去毒框架，该框架首先根据毒性的表现形式对文本进行分类，进而根据文本毒性表现形式的不同采用不同的处理方式，我们还引入了大语言模型（LLM）构建歧视词词典。实验表明，本文提出的模型能有效地处理中文文本中的性别歧视问题。

**关键词：** 文本风格迁移；性别歧视；大语言模型

# Research on detoxification of Chinese sexist texts based on text style transfer

**Jian Peng　Jiali Zuo*　Jingxuan Tan　Jianyi Wan　Mingwen Wang**
Email: {pengjian2022, zjl, tanjingxuan, mwwang}@jxnu.edu.cn, wanjianyi@aliyun.com

**Abstract**

There is a certain degree of sexist speech on online social media platforms, which hinders the health of the Internet and the development of social civilization. Text style transfer technology can reduce sexism in text, which has been studied in English and other languages. However, in the field of Chinese, due to the lack of data sets, there are few related studies. In addition, due to the rich semantic information and diverse language expressions in Chinese, the toxicity of sexist speech is manifested in various forms, and the existing methods mostly adopt a single text style transfer model, which is not effective. Therefore, this paper proposes a framework for detoxifying Chinese sexist texts based on text style transfer, which first classifies texts according to their manifestations of toxicity, and then adopts different processing methods according to different manifestations of text toxicity. We also introduce a large language model to construct a dictionary of discriminatory words. Experiments show that the model proposed in this paper can effectively deal with gender discrimination in Chinese texts.

**Keywords:** Text style transfer , Sexism , Large language model

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版
**基金项目：** 国家自然科学基金(61866018,62266023)
**通讯作者：** 左家莉

## 1 引言

相较于现实生活空间，网络空间具有其内在的诸多特殊性，如虚拟性、匿名性、间接性、开放性、流动性等，这些特性决定了网络空间是一个身份遮蔽的场所 (李建华, 2021)。基于这些特性，在网络社交媒体上，用户可以匿名地对特定群体或个人发布辱骂性言论或表现攻击性态度，这就导致网络上存在大量的仇恨言论，学术界通常将它定义为使用语言煽动暴力或促进对特定群体的暴力或仇恨，或基于特定形式的攻击、侮辱、贬低群体成员 (Fortuna and Nunes, 2018)，是一种典型的有毒言论。

性别歧视言论是一种常见的仇恨言论，目前被认为是中国社交网络恶化的原因之一 (Waseem and Hovy, 2016)。Jiang等人 (2022)认为如果文本出现以下任何一种情况，则其可被视为性别歧视：1）使用性别歧视语言攻击或侮辱性别群体或个人。2）煽动基于性别的暴力或宣扬性别歧视的仇恨，但不直接使用性虐待语言。3）虐待那些攻击或对性别群体持负面态度的人。4）表示支持有问题的事件或表达性侵犯、性取向和性骚扰的意图。5）通过描述身体吸引力、过度简化形象或表达男性优于女性，对性别群体产生负面的刻板印象。6）以讽刺或隐晦的方式表达潜在的性别偏见。其它文本被认为是非性别歧视的，这包括对性相关事件或现象的中性描述或证词。性别歧视言论的存在污染了网络环境，会对特定群体造成不良影响，这一类言论我们都归为毒性言论。Tokpo等人 (2022)认为使用文本风格迁移（Text Style Transfer, TST）技术可以实现文本去毒。

文本风格迁移是自然语言处理领域中的一项重要任务 (Nouri, 2022)。它旨在将原始句子的风格转换为目标风格并保留原始句子的内容。句子的风格包括但不限于情感、礼貌和幽默，句子毒性通常被视为一种风格 (Santos et al., 2018)。由于数据的标注成本较高，导致文本风格迁移任务缺少平行数据集，因此大多数工作都是以无监督的方式进行。总体而言，文本风格迁移的方法可以分为两种 (Han et al., 2023)：1）编辑式方法考虑到句子的风格通常由特定短语或词汇决定，所以该方法通过只改变具有风格特征的词汇来实现文本风格迁移。2）生成式方法一般基于编码器-解码器模型架构，通过编码器将文本映射到潜在空间得到文本的潜在内容表示，再使用解码器对其解码得到目标风格文本。编辑式方法仅替换风格词或短语，因此内容保留度高，生成式方法能够对整个文本进行重写，所以文本流畅度较高。

有研究表明，毒性的表现形式可以归纳为显性和隐性 (Lu et al., 2023)，显性文本通常包含明显冒犯、煽动或偏见，包含强烈的攻击性、侮辱性词汇；而隐性文本的毒性则表现得较为隐晦，它通常不包含直接的侮辱性词汇，如表1所示，中文性别歧视文本常通过谐音词、“字形梗”、明褒暗贬等隐晦的方式表现。我们发现使用单一的文本风格迁移模型来处理这些不同表现形式的文本效果不佳。显性毒性的文本具有毒性词外显的特点，更适合使用编辑式模型；而对于隐性毒性的文本，生成式模型则能够对整个文本进行重写。

本文主要对中文领域的性别歧视文本去毒进行研究，我们重构了一个包含性别歧视的中文文本数据集，并构建了一个基于文本风格迁移的中文性别歧视文本去毒框架，首先该框架根据毒性的表现形式将文本分为显性或者隐性，对于显性的毒性文本该框架采用编辑策略，对于隐性的毒性文本该框架采用改写策略，为了提高风格迁移强度，该框架引入大语言模型（LLM）构建歧视词词典用于文本编辑。

| | |
|---|---|
| 显性 | 女人真是心狠 |
| | 女司机是马路杀手 |
| | 楼主优秀，男人怂货，怂货都虚伪 |
| 隐性 | 来看看蝈女的素质 |
| | 珍爱生命，远离幕刃 |
| | 女生嘛，做点家务怎么了 |

Table 1: 两种表达的歧视文本示例

## 2 相关工作

### 2.1 文本中的性别歧视研究

在性别歧视相关数据集和资源方面，大多数研究都是针对英文进行的 (Lu et al., 2023)，中文相关语料的匮乏阻碍了相关领域的研究和发展。Jiang等人 (2022)提出了第一个中文性别


(c) 2024 中国中文信息学会计算语言学专业委员会

歧视数据集——新浪微博性别歧视（SWSR）数据集，这个数据集由新浪微博上与性别歧视相关的微博和评论组成。此外，与英文相比，中文在性别歧视的表达上更加丰富，它更容易出现隐性毒性的文本，而现有数据集缺少对毒性类型的描述和细粒度标注 (Lu et al., 2023)。Lu等人 (2023)为了填补这一空白，提出了MONITOR TOXIC FRAME用于分析毒性类型和表达，并提出了一个细粒度标注的数据集，这个数据集包含了显性和隐性毒性的性别歧视数据。

### 2.2 文本风格迁移

文本风格迁移任务的目标是改变文本风格的同时，保留与风格无关的内容文本。由于该任务普遍缺乏平行数据，所以大多数文本风格迁移任务都是采用无监督学习的方法，该方法大致可以分为编辑式和生成式方法。

编辑式方法依赖于风格属性词与内容词的表面级分离 (Li et al., 2018)，具体做法是区分文本中的风格属性词和内容词，在生成时用目标风格词替换原始句子中的风格词，从而达到改变文本风格的目的。在这类方法中，识别文本中与风格相关的词是至关重要。删除-检索-生成（DRG）文本风格迁移模型 (Li et al., 2018)使用基于词频统计的方式识别输入文本的风格属性词并删除，只保留与风格无关的内容词，再检索目标风格词，最后将内容词和目标风格词输入进RNN生成目标风格文本。此外，其它风格迁移模型利用MLM（掩码语言模型）识别文本中的风格属性词，例如，Shi等人 (2023)最近的一项工作提出了AMR-TST，该模型利用RoBERTa模型对输入文本中的词进行评分，这些评分反映了词对文本风格贡献的程度，分数越高，当前词是风格词的概率越大。Wu等人 (2019)使用BERT对风格词进行掩码，最后将掩码标记替换为目标风格词，从而实现风格风格迁移。Dale等人 (2021)对上述模型进行了改进，提出了CondBERT模型，它使用毒性分类器识别风格属性词，在替换风格属性词时采用内容保存的启发式方法保留被替换词的语义，以选择合适的目标风格词。编辑式方法的优势在于内容保留度较高，但此类方法生成的文本在流畅度方面表现较差。

生成式方法一般通过学习内容和风格的潜在表示，隐式过滤文本中与风格相关的词，然后生成目标风格的文本。这种方法一般基于Encoder-Decoder（编码器-解码器）架构，由Encoder对输入文本编码得到潜在内容表示，再由Decoder对潜在表示解码生成目标风格文本。John等人 (2019)将潜在表示分解为内容表示和风格表示，再用目标风格表示替换原风格表示，最后由生成模型生成目标风格文本；Hu等人 (2017)强制Encoder表示与风格无关的内容，再由判别器指导生成模型生成目标风格文本；Wang 等人 (2019)采用多任务学习的方法实现文本的风格迁移；Dale等人 (2021)基于GeDi (krause et al., 2021)提出了ParaGeDi，它由生成模型和判别模型组成，对输入文本进行重写，在重写的过程中受到判别模型的监督，从而生成目标风格文本。这类方法的优点在于流畅度较高，但由于其改变了文本中的大部分内容，所以内容保留度较差。

### 2.3 大语言模型

大语言模型(Large Language Model，LLM)是指包含数千亿(或更多)参数的Transformer语言模型 (Shanahan et al., 2024)。这些模型是在海量文本数据上训练的，如GPT4 (OpenAI et al., 2023)、ChatGLM (Zeng et al., 2022)、BaiChuan (Yang et al., 2022)等。Ye等人 (2022)的研究展示了LLM在使用zero-shot提示方面的能力，仅需简单的提示如“The movie review with positive sentiment is: ”，就能有效地引导LLM生成平行语料。并且，思维链(Chain-of-Thought，CoT)已被证明可以提高LLM在NLP任务上的性能 (Shaikh et al., 2023)。Wei等人 (2022)使用CoT提示的方式使LLM在算术、常识和符号推理等任务上的性能有所提升，证明了CoT提示可以激发LLM的推理能力，进而满足人类更复杂的任务需求。我们发现使用大语言用于检测风格属性词会更加全面。

## 3 方法

本文提出了一个新的文本去毒框架，该框架结合了编辑和改写策略，使其可以同时处理两种不同毒性表现形式的毒性文本，对于显性的毒性文本该框架采用编辑策略，对于隐性的毒性文本该框架采用改写策略。具体来说：它包含类别判别模块、改写模块、编辑模块和词典构建模块。该框架如图1所示，类别判别模块根据毒性的表现形式对输入文本进行分类，依据分类结果，改写模块调用生成式模型处理隐性类的毒性文本，编辑模块则对显性类的毒性文本进行编


(c) 2024 中国中文信息学会计算语言学专业委员会

辑处理，为了提高风格迁移强度，我们在此处使用了词典构建模块，该模块引入大语言模型构建歧视词词典用于文本编辑。该框架针对中文文本进行了优化，还在毒性内容检测方面提出了新的解决策略。

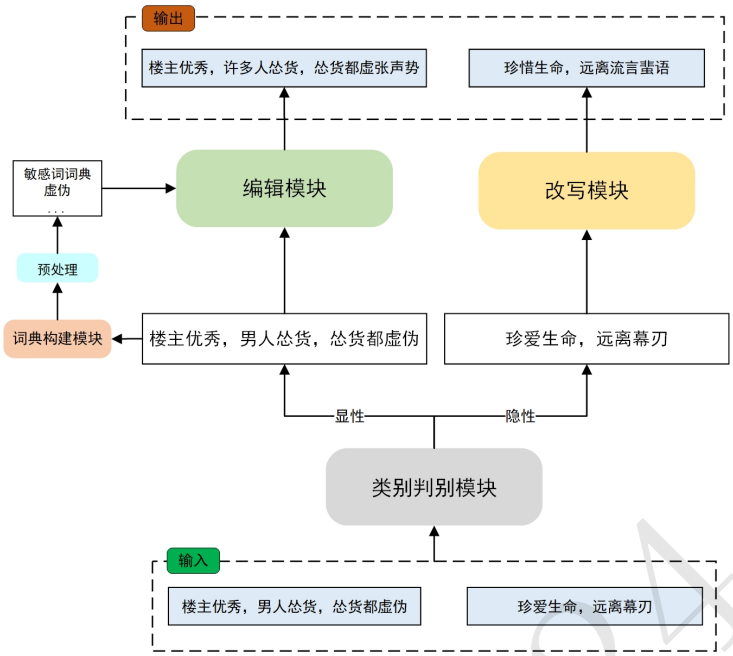



Figure 1: 模型整体架构

### 3.1 任务描述

我们将无监督文本风格迁移表述为：$S = \{s_{src}, s_{tgt}\}$分别表示原风格和目标风格，在本文中，原风格和目标风格分别代表歧视风格和中性风格。我们把原风格的句子表示为$X = \{x_1 x_2 \cdots x_n\}$，其中$x_1 x_2 \cdots x_n$表示句子中的词；把目标风格的句子表示为的句子$Y = \{y_1 y_2 \cdots y_m\}$，其中$y_1 y_2 ... y_m$表示句子中的词。任务目标是将风格为$s_{src}$的原始文本$X$转换为风格为$s_{tgt}$的目标文本$\hat{Y} = \{\hat{y}_1 \hat{y}_2 \cdots \hat{y}_t\}$，同时尽可能地保留原始文本$X$中的内容。

### 3.2 类别判别模块

预训练语言模型BERT (Devlin et al., 2019)在诸多任务中取得了显著成果，其中之一便是文本分类，本研究需要根据毒性的表现形式将输入文本分类为显性或者隐性。我们微调预训练语言模型bert-base-chinese[0]作为文本分类模型，模型由输入层、编码层和分类输出层构成，如图2所示：

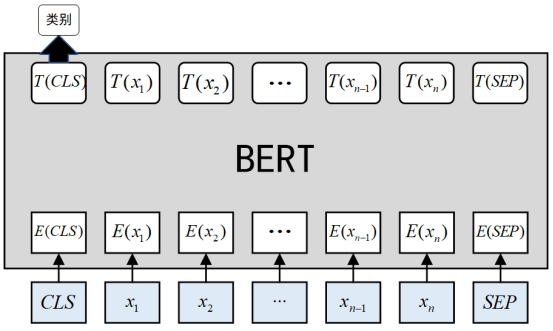



Figure 2: 基于BERT的文本分类模型

[0]https://huggingface.co/google-bert/bert-base-chinese


(c) 2024 中国中文信息学会计算语言学专业委员会

模型输入层使用词向量矩阵、块向量矩阵和位置向量矩阵对$X$进行映射，得到输入表示$v$：

$$X = [CLS]x_1x_2\cdots x_n[SEP] \tag{1}$$

$$v = InputRepresentation(X) \tag{2}$$

在BERT编码层中，输入$v$经过多层Transformer的编码，借助自注意力机制学习句子中每个词之间的语义关系，并最终得到句子的上下文语义表示$h$：

$$h = BERT(v) \tag{3}$$

由于BERT预训练阶段的NSP任务使用了[CLS]位预测，通常在文本分类任务中也是用同样的方法预测文本类型。模型使用[CLS]位对应的隐藏层表示$h_0$预测输入文本对应的类别标签。由下式表示：

$$P = Softmax(h_0W^0 + b^0) \tag{4}$$

式中，$W^0$表示全连接层的权重，$b^0$表示全连接层的偏置。

### 3.3 改写模块

改写模块使用了ParaGeDi模型，它是一个生成式模型，用于重写隐性毒性的文本。本文使用的ParaGeDi模型在原模型基础上进行了改进。具体而言，ParaGeDi为了保留输入文本的基本内容，用释义模型替换语言模型，使得ParaGeDi可以重写特定风格的文本。为此，该任务使用重构的中文语料库重新训练了生成式语言模型BART（chinese-bart-paraphrase[1]）作为中文释义器，并微调GPT2(gpt2-base-chinese[2])得到风格判别模型。ParaGeDi对以下概率建模：

$$P(\hat{y}_t|\hat{y}_{<t}, X, s_{tgt}) \propto P_{LM}(\hat{y}_t|\hat{y}_{<t}, X)P_D(s_{tgt}|\hat{y}_t, \hat{y}_{<t}) \tag{5}$$

其中$X = \{x_1x_2\cdots x_n\}$表示输入文本，$s_{tgt}$表示目标风格。模型根据原始句子、目标风格属性和$t$时刻前生成的文本预测$t$时刻的输出。$t$时刻的输出概率由两项决定，第一项是释义模型$P_{LM}$产生的，第二项使用贝叶斯规则与类条件语言模型$P_{CC}$求解得到，如下述公式所示：

$$P_D(s_{tgt}|\hat{y}_t, \hat{y}_{<t}) \propto P(s_{tgt})P_{CC}(\hat{y}_t, \hat{y}_{<t}|s_{tgt}) \tag{6}$$

因此，ParaGeDi的训练损失由两个损失线性组合而成：LM训练使用生成损失$L_G$，以及风格判别损失$L_D$，各类损失函数如下：

$$L_G = -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i}\sum_{t=1}^{T_i} logP\left(y_t^{(i)}|y_{t<t}^{(i)}, s_i\right) \tag{7}$$

$$L_D = -\frac{1}{N}\sum_{i=1}^{N} logP\left(s_i|y_{1:T_i}^{(i)}\right) \tag{8}$$

$$L_{ParaGeDi} = \lambda L_D + (1-\lambda)L_G \tag{9}$$

该模型如图3所示：

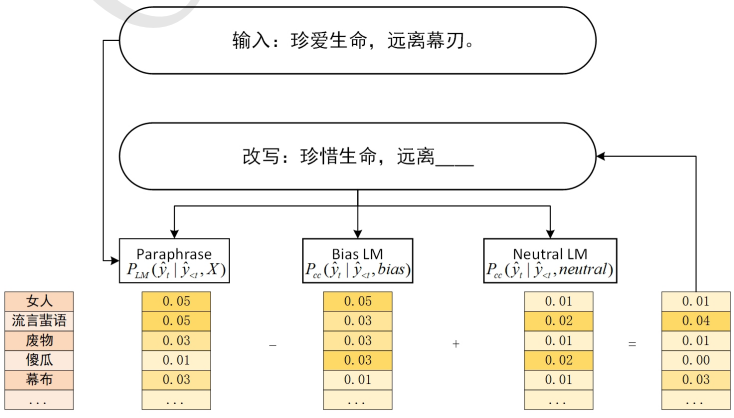



Figure 3: ParaGeDi模型概述

[1]https://huggingface.co/figurative-nlp/chinese-bart-paraphrase
[2]https://huggingface.co/ckiplab/gpt2-base-chinese


(c) 2024 中国中文信息学会计算语言学专业委员会

### 3.4 编辑模块

如图4所示，本文在编辑模块使用CondBERT模型。具体而言，通过训练一个毒性分类器，在训练过程中分类器将句子的单个词作为特征，为每个特征（词）分配一个权重作为毒性分数$s$。当毒性分数大于阈值$t = \max(t_{min}, \max(s_1, s_2, \cdots, s_n)/2)$的词定义为有毒单词，其中，$s_1, s_2, \cdots, s_n$代表句子中所有单词的分数，$t_{min}$设置为0.2为最小毒性分数。然后将有毒单词用[MASK]替代，最后用BERT将掩码标记替换为目标风格词。此外，为了保留被替换词的语义，该模型采用了内容保存的启发式方法保留原风格词的语义信息，根据候选词与替换词的语义相似度进行重排序以选择更合适的候选词。

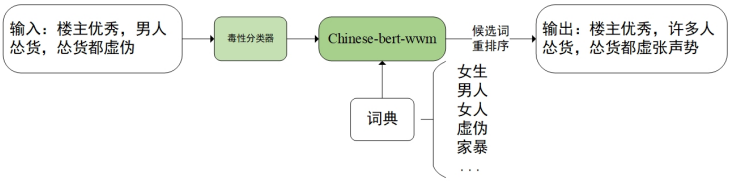



Figure 4: CondBERT模型概述

在实际应用中，由于训练语料的局限性，毒性分类器通常不能识别出句子中所有的歧视词，而且大语言模型在各项语言理解任务上展现出了卓越性能，所以我们使用大语言模型来构建歧视词词典。已有研究发现，CoT提示可以引导模型进行推理，使模型更全面地考虑问题，从而提供更完整的答案 (Wei et al., 2022)。不使用CoT提示会导致模型错误识别或者漏识别关键歧视词。如图5所示，我们基于大语言模型（BaiChuan2-13b[3]）使用思维链提示对文本中关键歧视词进行识别，对模型的输出结果进行提取以构建歧视词词典。相较于之前的方法，使用大模型构建的歧视词词典识别效果更好，能提升模型的去毒能力。

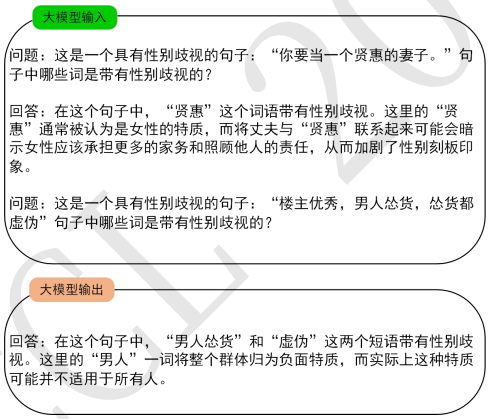



Figure 5: 思维链推理过程

经典的BERT模型使用WordPiece算法，它会把一个完整的词切分为若干个子词，在生成训练样本时，这些子词将被随机掩码。而在中文环境下，相对于单个字，完整的词通常包含更多的语义信息，所以我们将子词掩码替换为全词掩码（Whole Word Masking）Chinese-BERT-wwm[4]。在该设置下，若一个词的部分子词被掩码，则同属该词的其它部分也会被掩码，从而可以实现对包含多个字的中文词语的掩码。

## 4 实验

### 4.1 数据集

本文重构的数据集主要来源于三部分：（1）Tokpo等人 (2022)通过筛选Jigaw dataset构建了一个性别偏见数据集[5]，我们通过调用翻译API的方式将英文文本翻译为中文。（2）

[3]https://github.com/baichuan-inc/Baichuan2
[4]https://github.com/ymcui/Chinese-BERT-wwm
[5]https://www.kaggle.com/c/Jigsaw-unintended-bias-in-toxicity-classification/data


(c) 2024 中国中文信息学会计算语言学专业委员会

从Deng等人 (2022)提出的中文冒犯性语言检测数据集（COLDataset[6]）中筛选主题为性别的文本。（3）Jiang等人 (2022)提出了新浪微博性别歧视评论（SWSR[7]）数据集，我们从中选取微博评论文本。数据集大小如表2所示：

| | 歧视文本 | 中性文本 | 总计 |
|---|---|---|---|
| 训练集 | 95732 | 88889 | 184621 |
| 验证集 | 1000 | 1000 | 2000 |
| 测试集 | 500 | 500 | 1000 |
| 总计 | 97232 | 90389 | 187621 |

Table 2: 数据集统计

## 4.2 评估

### 4.2.1 自动评估

在文本风格迁移任务实现的目标包括：生成文本要符合目标风格，保留输入文本中与风格无关的内容，并具有人类写作特征 (Mir et al., 2019)。本文使用文本风格迁移任务中常用的自动评估指标，从以下几个方面对模型的性能进行评估：

**风格迁移强度**：该指标用于评估模型成功将性别歧视文本转换为中性文本的百分比。为此，我们使用表2所示的数据集，对bert-base-chinese进行微调。将微调后的BERT模型作为验证模型，利用其预测结果统计最终的ACC 得分。

**内容保留度**：BLEU (Papineni et al., 2002)分数是常见的文本生成评估指标，它计算候选文本于参考文本之间的n-gram重叠，未考虑到文本间的语义信息。Narasimhan等人 (2022)提出，仅依靠BLEU分数不足以衡量生成文本与参考文本的相关性。因此，本文引入了BertScore (Zhang et al., 2019)进行内容保留的评估，这是一种基于预训练BERT上下文嵌入的文本生成评估指标。BertScore将两个句子的相似度计算为它们的词嵌入之间的余弦相似度之和。BertScore分数越高，说明生成文本中保留了更多的内容。

**流畅度**：该指标用于衡量生成文本的流畅度，本文使用在中文数据上微调过的gpt2-chinese-cluecorpussmall[8]计算生成文本的困惑度（PPL）。指标越小说明文本的流畅度越高。

**几何平均数**：为了体现模型的综合性能，文本将ACC、BLEU、BERTScore和$\frac{1}{\ln PPL}$的几何平均值作为综合评估指标，用GM表示。

### 4.2.2 人工评估

为了更加灵活和全面地评估模型生成的文本，我们进行了人工评估。具体方法是，从各模型生成的文本中随机选择100条，并提供相应的原始文本，要求评估人员从风格迁移强度（ACC）、流畅度（FL）、内容保留度（SIM）这三个维度进行评分。评分标准是从1（非常差）到5（非常好）的范围内打分。我们取七名评估人员（其中四名男性，三名女性）的平均分作为评估结果。

## 4.3 基线模型

我们与四个现有模型进行比较。1）**DRG**：DRG通过计算每个词的相对频率来识别输入句子的风格属性词。然后模型删除原句子中的风格属性词，其余文本则作为无风格的内容文本。接着DRG从目标风格语料库中检索与内容文本语义相似的句子，再从该句子中提取目标风格属性词。最后DRG结合内容文本和提取出的目标风格属性词生成目标风格文本 (Li et al., 2018)。2）**Style Transformer**：Style Transformer模型遵循标准的编码器-解码器体系结构，将输入句子映射为潜在表示，由解码器对前者表示解码。为了控制解码器生成文本的风格，作者使用判别模型对模型的生成过程进行监督 (Dai et al., 2019)。3）**TWR**：Yang等人 (2022)认为没有必要识别风格属性词并且使用纯粹的生成方法改写与风格无关的词是多余的。所以TWR模型首先通过条件随机场（CRF）标记句子中的词。标记有四种，分别是[INS]（插

[6]https://github.com/thu-coai/COLDataset
[7]https://zenodo.org/records/4773875
[8]https://huggingface.co/uer/gpt2-distil-chinese-cluecorpussmall


(c) 2024 中国中文信息学会计算语言学专业委员会

入）、[SUB]（替换）、[DEL]（删除）和[KEEP]（保留）。在一个句子被标记后，模型根据与句子中的词相对应的操作来生成目标风格的句子。4）**Bias Mitigation**：Bias Mitigation 是一个用来缓解句子中蕴含对女性偏见的模型。它使用LIME检测风格属性词，并对其掩码，再对掩码后的句子生成一系列词嵌入表示。然后用潜在内容编码器对原始句子（未掩码）编码为潜在内容表示。接着根据词嵌入表示和潜在内容表示生成新的词嵌入表示，最后用解码器对其解码生成目标风格句子 (Tokpo and Calders, 2022)。

### 4.4 实验设置

对于ParaGeDi，我们使用重新训练的chinese-bart-paraphrase作为释义器，并且将微调后的gpt2-base-chinese作为判别器。在生成目标风格文本时采用集束搜索（beam search）的方式，我们将beam size设置为10。temperature用于控制生成文本的随机性和多样性，较高时会更平均地分配概率给各个词，这会导致生成的文本更具随机性和多样性；当temperature趋近于0时，模型更倾向于选择概率较高的词，在本次任务中，temperature设置为1。repetition_penalty的目标是惩罚模型生成重复的词，以减少生成文本的重复性，我们将repetition_penalty设置为3。对于CondBERT，风格的控制通过分类器来完成，我们未进行微调。contrast_penalty是为了惩罚生成语义相似的词，本次任务设置为0，不对生成语义相似的词进行惩罚；n_top决定候选词的个数，模型通过对候选词重排序选择与被替换词语义最相似的词作为目标风格词，此次任务设置为10，其余超参数则遵循Dale等人 (2021)的设置。

## 5 结果与分析

### 5.1 自动评估结果

所有测试模型的性能如表3所示，大多数基线模型难以在流畅度、风格迁移强度与内容保留度这三个评估指标上实现良好的平衡。例如$\text{DRG}_{\text{RetireveOnly}}$，它拥有较好的流畅度和风格迁移强度，但其内容保留度较差。这是因为在RetrieveOnly设置下，DRG模型仅返回检索到的目标文本。这可以尽可能保证生成一个具有目标风格的文本，但内容可能与输入文本不同。相比之下，$\text{DRG}_{\text{deleteOnly}}$在内容保留度上表现更佳，因为在这种设置下，它只删除风格属性词，最大程度地保留了非风格内容。然而，这种设置下的文本因删除风格词可能出现语法错误，导致流畅度下降。TWR模型的ACC分数虽高，但因删除了大量原始文本词汇，包括影响风格判别的歧视词和内容词，导致其BLEU分数较低。Bias Mitigation模型的流畅度较差，这是因为模型在生成文本的过程中，会产生一些英文文本，这就导致文本的语法不规范，可读性较差。本研究提出的模型在这三个目标之间能够保持良好的平衡，尤其在内容保留度方面表现最佳，并且在其它指标上都有较强的竞争力。这可能源于我们考虑到文本风格迁移模型的特性，有针对性地选择输入文本进行编辑或重写,充分发挥了编辑式和生成式方法的优势。

| Model | ACC | BLEU | BertScore | PPL | GM |
|---|---|---|---|---|---|
| $\text{DRG}_{\text{deleteOnly}}$ | 17.8 | 63.8 | 0.895 | 494.2 | 3.57 |
| $\text{DRG}_{\text{delete\&Retireve}}$ | 20.2 | 64.7 | 0.893 | 565.6 | 3.64 |
| $\text{DRG}_{\text{RetrieveOnly}}$ | 69.4 | 0.38 | 0.557 | **69.2** | 1.36 |
| TWR | **77.2** | 0.11 | 0.640 | 724.2 | 0.95 |
| BiasMitigation | 20.4 | 60.2 | 0.890 | 1994.5 | 3.45 |
| $\text{StyleTransformer}_{\text{Cond}}$ | 34.8 | 44.9 | 0.796 | 248.9 | 3.87 |
| $\text{StyleTransformer}_{\text{Mutil}}$ | 37.8 | 38.3 | 0.794 | 147.4 | 3.89 |
| Ours | 31.4 | **79.4** | **0.931** | 167.4 | **4.64** |

Table 3: 自动评估结果

### 5.2 人工评估结果

在自动评估中，我们选取了表现最优的两个基线模型与本研究提出的模型进行比较。评估结果如表4所展示，与自动评估的结果大体一致。本研究提出的方法在风格迁移强度、流畅度和内容保留度三个指标上均取得了较高的评分。


(c) 2024 中国中文信息学会计算语言学专业委员会

| Model | ACC | FL | SIM |
|---|---|---|---|
| StyleTransformer$_{\text{Cond}}$ | 2.5 | 2.2 | 1.9 |
| StyleTransformer$_{\text{Mutil}}$ | 2.4 | 2.3 | 2.0 |
| Ours | **2.9** | **3.1** | **4.0** |

Table 4: 人工评估结果

### 5.3 消融实验

为了研究不同模块对整体的性能，我们进一步对文本提出的模型进行了消融实验。结果如表5所示：

**w/o All.**：为了更好地判断风格判别模块和外部知识词典对模型性能的影响，我们删除了风格判别模块，去除了歧视词词典，单独训练ParaGeDi和CondBERT进行测试。

**w/o Dic.**：为了探索词典对模型性能的影响，我们去除词典，并使用相同的模型参数重新进行训练。

**Stand.**：不删除任何模块。

| | Model | ACC | BertScore | PPL |
|---|---|---|---|---|
| w/o All. | ParaGeDi | 26.6 | 0.869 | **164.0** |
| | CondBERT | 15.8 | **0.946** | 194.0 |
| w/o Dic. | Ours | 18.6 | 0.939 | 170.7 |
| Stand. | Ours | **31.4** | 0.931 | 167.4 |

Table 5: 消融实验结果

从实验结果来看，词典与类别判别模块有助于提高模型的综合性能。在明显提高风格迁移强度的同时，内容保存度和流畅度都具有较强的竞争力。

### 5.4 参数选择

我们进一步分析了候选词数量对模型内容保留度的影响。具体来说，这可以通过调整n_top的数值来观察，n_top的值表示候选词的个数，模型根据候选词与被替换词的相似度进行重排序，并选择相似度最高的候选词作为目标风格词。在保持其它参数不变的情况下，我们调整n_top的大小进行实验。如图6所示，随着候选词数量的增加，BLEU分数呈上升的趋势，这表明增加候选词的数量有助于提高模型的内容保留度，也证明模型对候选词进行重排序是有效的。

集束搜索（beam search）中beam的大小会影响生成文本的流畅度(Dale et al., 2021)，我们固定其它参数不变，通过改变束的大小来观察文本困惑度（PPL）的变化。如图7所示，当束的大小从1增加到5时，文本的困惑度显著下降，之后略有回升，但随着束数量的进一步增加，困惑度又逐渐降低。总体而言，困惑度的下降趋势表明文本的流畅度得到了提升。

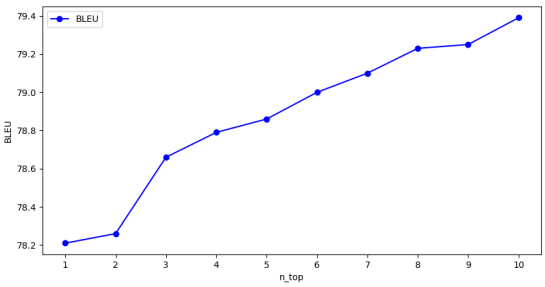



Figure 6: 候选词的个数对BLEU的影响

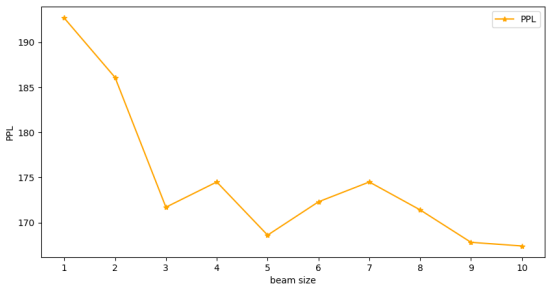



Figure 7: beam的数量对PPL的影响


(c) 2024 中国中文信息学会计算语言学专业委员会

### 5.5 类别判别模块性能对比

如图8所示，我们在类别判别模块对比了bert-base-chinese和roberta-base两个文本分类模型，最终选择Accuracy数值更高的模型。

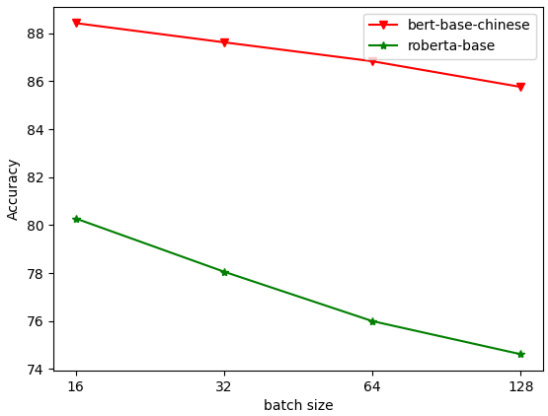



Figure 8: 类别判别模块性能

### 5.6 大模型检测结果

总体而言，大模型已经能够很好地识别出文本中的关键歧视词。但是，大模型很难识别谐音词和所谓的“字形梗”。同时，大模型也不擅长处理包含多个歧视词的文本。部分示例如表6所示。

| 输入 | 输出 |
|---|---|
| 这是一个具有性别歧视的句子：“快进到**玉玉症**退网。”句子中哪些词是带有性别歧视的? | 这个句子中没有明显的性别歧视词汇。然而，句子的含义可能涉及到对女性的贬低和侮辱。在这种情况下，我们可以认为整个句子都带有性别歧视。 |
| 这是一个具有性别歧视的句子：“珍爱生命，远离**幕刃**”句子中哪些词是带有性别歧视的? | 在这个句子中，没有明显的性别歧视词汇。然而，这句话可能被认为是性别歧视，因为它暗示了女性可能会受到伤害或危险。为了避免引起误解，可以考虑使用更中性或更具包容性的表达方式。 |
| 这是一个具有性别歧视的句子：“普信**幕刃**罢了。”句子中哪些词是带有性别歧视的? | 这个句子中的“普信幕刃”并没有明确的性别歧视含义，但其中的“罢了”可能带有贬义和轻视的意味。然而，这句话的整体含义并不明确，很难确定其中具体的性别歧视成分。 |
| 这是一个具有性别歧视的句子：“立即开启**印度模式**省得**她们**作。”句子中哪些词是带有性别歧视的? | 这个句子中的“**她们**”这个词带有了性别歧视，因为在这里它被用来泛指女性，而忽略了男性也可能参与作的情况。这种表达方式可能被认为是将女性归为容易作闹的一类，从而加剧了性别刻板印象。 |
| 这是一个具有性别歧视的句子：“**女性**和**幕刃**是两种生物。”句子中哪些词是带有性别歧视的? | 在这个句子中，“**女性**”这个词是带有性别歧视的。因为这句话将女性与另一种生物（在这里指的是男性）进行对比，暗示了女性与男性之间存在某种差异或不平等。而“幕刃”这个词本身没有性别歧视的含义。 |
| 这是一个具有性别歧视的句子：“我185，但这**幕刃**是**鲨bee**”句子中哪些词是带有性别歧视的? | 这个句子的确存在性别歧视。其中包含性别歧视的词是“**鲨bee**”，这里的“bee”是对女性的贬义称呼。 |

Table 6: 大模型检测部分错误示例

我们认为，上述问题的产生可能是由于训练大型语言模型所使用的数据集是标准的、滞后的。然而，网络流行语言的发展速度极快，其中一些谐音和“字形梗”是基于特定社会背景而产生的，并随着网络流行语言的演变而不断涌现。因此，大模型可能难以理解这些新兴词汇的具体含义，从而导致无法准确识别。


(c) 2024 中国中文信息学会计算语言学专业委员会

### 5.7 质量检验

本文提出的模型与基线生成的文本示例如表7所示。对于基线模型$\text{DRG}_{\text{retireveonly}}$，示例显示，它的生成较为完整，但是生成内容与原始文本不一致，这与表3中模型的自动评估结果相符，即流畅度得分高，但是文本保留度较差。我们还可以观察到，StyleTransformer模型生成的文本可读性不高，会在一句完整的句子后面继续生成一些与内容无关的词，这可以解释它生成的文本内容保留度低。对于本文提出的模型，它生成的文本能够很好的减轻文本性别歧视，并且能保留大部分的原文本内容。此外，与其它基线模型生成的文本相比，没有出现明显的语法错误，适合人类阅读，这对于在现实场景中应用TST模型是可靠的。

| **示例1** | |
|---|---|
| Source | 珍爱生命，远离幕刃。 |
| $\text{DRG}_{\text{deleteOnly}}$ | 做好生命，远离另一方。 |
| $\text{DRG}_{\text{delete\&retrieve}}$ | 做些他们，远离拳头。 |
| $\text{DRG}_{\text{retrieveOnly}}$ | 我也觉得有性别歧视的女性是处，但这很正常。 |
| TWR | 生，。生。 |
| BiasMitigation | 珍爱生命，远离幕刃。 |
| $\text{StyleTransfomer}_{\text{cond}}$ | 生命，远离幕刃。 |
| $\text{StyleTransfomer}_{\text{mutil}}$ | 生命，远离战场 |
| Ours | 珍惜生命,远离流言蜚语。 |
| **示例2** | |
| Source | 女司机是马路杀手。 |
| $\text{DRG}_{\text{deleteOnly}}$ | 佩里强年轻小事。 |
| $\text{DRG}_{\text{delete\&retrieve}}$ | 日期是施虐者。 |
| $\text{DRG}_{\text{retrieveOnly}}$ | 我觉得是的，我觉得我是性别歧视，但是我不喜欢她 |
| TWR | 女。 |
| BiasMitigation | 女司机是马路杀手 |
| $\text{StyleTransfomer}_{\text{cond}}$ | 一千年是马路杀手。一千年是马路。有意无意。有意无意马路 |
| $\text{StyleTransfomer}_{\text{mutil}}$ | 这特是马路政治家。这特是别人是马路。的。 |
| Ours | 驾驶员是马路杀手。 |

Table 7: 模型生成文本示例

## 6 总结与展望

本文提出了一种新的框架，旨在减轻文本蕴含的性别歧视问题。与以往不同的是，该框架结合了编辑式和生成式文本风格迁移方法，并引入大语言模型以检测文本中的毒性跨度，将其提取构成敏感词词典。随后通过判别模型将歧视文本的表达类型分为显性或隐性。对于显性毒性文本采用编辑式方法；而对于隐性毒性文本，则使用生成式方法。为了使模型更好地适应中文环境，本文使用中文数据对生成模型进行了重新训练，并对判别模型进行了微调。实验结果表明，在相同数据集上，本文所提出的模型优于先前的文本风格迁移模型。此外，采用Chain-of-Thought（CoT）提示的大语言模型在一定程度上能够有效地识别文本中的有毒词汇，然而仍需进一步探究不同提示对大语言模型性能的影响。

在未来工作中，对于如何处理隐性毒性文本，我们将考虑引入上下文的情感信息与背景知识。此外，我们将进一步构建更细粒度的数据集展开研究，并探究大模型在文本去毒任务上的有效性。

## 参考文献

David Dale, Anton Voronov, Daryna Dementieva, Varvara Logacheva, Olga Kozlova, Nikita Semenov, Alexander Panchenko. 2021. Text Detoxification using Large Pre-trained Neural Models. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7979-7996.

Jiawen Deng, Jingyan Zhou, Hao Sun, Chujie Zheng, Fei Mi, Helen Meng, Minlie Huang. 2022. COLD:

A Benchmark for Chinese Offensive Language Detection. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11580-11599.

Ning Dai, Jianze Liang, Xipeng Qiu, Xuanjing Huang. 2019. Style Transformer: Unpaired Text Style Transfer without Disentangled Latent Representation. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5997–6007.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171-4186.

Paula Fortuna, Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. In *ACM Computing Surveys (CSUR)*, pages 1-30.

Jingxuan Han, Quan Wang, Licheng Zhang, Weidong Chen, Yan Song, Zhendong Mao. 2023. Text Style Transfer with Contrastive Transfer Pattern Mining. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7914-7927.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, Eric P. Xing. 2017. Toward controlled generation of text. *International conference on machine learning. PMLR*, pages 1587-1596.

Aiqi Jiang, Xiaohan Yang, Yang Liu, Arkaitz Zubiaga. 2022. SWSR: A Chinese dataset and lexicon for online sexism detection. *Online Social Networks and Media*,27(2022):100-182.

Vineet John, Lili Mou, Hareesh Bahuleyan, Olga Vechtomova. 2019. Disentangled Representation Learning for Non-Parallel Text Style Transfer. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 424-434.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, Nazneen Fatema Rajani. 2021. GeDi: Generative Discriminator Guided Sequence Generation. *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4929-4952.

Junyu Lu, Bo Xu, Xiaokun Zhang, Changrong Min, Liang Yang, Hongfei Lin. 2023. Facilitating fine-grained detection of chinese toxic language: Hierarchical taxonomy, resources, and benchmarks. *arXiv preprint arXiv:2305.04446 (2023).*

Juncen Li, Robin Jia, He He, Percy Liang. 2018. Delete, Retrieve, Generate: a Simple Approach to Sentiment and Style Transfer. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1865-1874.

Remi Mir, Bjarke Felbo, Nick Obradovich, Iyad Rahwan. 2019. Evaluating Style Transfer for Text. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 495–504.

Nasim Nouri. 2022. Text Style Transfer via Optimal Transport. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2532-2541.

Sharan Narasimhan, Suvodip Dey, Maunendra Desarkar. 2022. Towards Robust and Semantically Organised Latent Representations for Unsupervised Text Style Transfer. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 456-474.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774 (2023).*

Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311-318.

Omar Shaikh, Hongxin Zhang, William Held, Michael Bernstein, Diyi Yang. 2023. On Second Thought, Let's Not Think Step by Step! Bias and Toxicity in Zero-Shot Reasoning. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4454-4470.


(c) 2024 中国中文信息学会计算语言学专业委员会

Murray Shanahan. 2024. Talking about large language models. *Communications of the ACM*, 67(2):68-79.

Yang Shuo. 2022. Tagging Without Rewriting: A Probabilistic Model for Unpaired Sentiment and Style Transfer. *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 293-303.

Cicero Nogueira dos Santos, Igor Melnyk, Inkit Padhi. 2018. Fighting Offensive Language on Social Media with Unsupervised Text Style Transfer. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 189-194.

Kaize Shi, Xueyao Sun, Li He, Dingxian Wang, Qing Li, Guandong Xu. 2023. AMR-TST: Abstract Meaning Representation-based Text Style Transfer. *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4231-4243.

Ewoenam Kwaku Tokpo, Toon Calders. 2022. Text Style Transfer for Bias Mitigation using Masked Language Modeling. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 163-171.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35(2022):24824-24837.

Zeerak Waseem, Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. *Proceedings of the NAACL Student Research Workshop*, pages 88-93.

Xing Wu, Tao Zhang, Liangjun Zang, Jizhong Han, Songlin Hu. 2019. "Mask and Infill": Applying Masked Language Model to Sentiment Transfer. *Proceedings of the Twenty-Eighth International Joint Con-ference on Artificial Intelligence, IJCAI-19*, pages 5271-5277.

Ke Wang, Hang Hua, Xiaojun Wan. 2019. Controllable unsupervised text attribute transfer via editing entangled latent representation. *Proceedings of the 33rd International Conference on Neural Information ProcessingSystems*, pages 11036-11046.

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, Lingpeng Kong. 2022. ZeroGen: Efficient Zero-shot Learning via Dataset Generation. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11653-11669.

Aiyuan Yang, Bin Xiao, Bingning Wang et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305 (2023)*.

Aohan Zeng, Xiao Liu, Zhengxiao Du et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414 (2022)*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675 (2019)*.

李建华. 2021. 网络空间道德建设中的自我伦理建构. 思想理论教育,(03):9-14.

CCL 2024

# 基于问题扩展的散文答案候选句抽取方法研究

**雷洋[1]，王素格[1,2,*]，李书琪[1]，王浩[1]**
1.山西大学计算机与信息技术学院，山西省 太原 030006
2.山西大学计算智能与中文信息处理教育部重点实验室，山西省 太原 030006
wsg@sxu.edu.cn

## 摘要

在散文阅读理解中，一方面问题的题干通常较为简洁、用词较为抽象，机器难以直接理解问题的含义和要求；另一方面，散文文章较长，答案候选句分散在文章的多个段落，给答案候选句的抽取任务带来巨大的挑战。因此，本文提出了一种基于问题扩展的散文答案候选句抽取方法。首先，利用大语言模型抽取文章中与问题题干相关的词，构建问题词扩展库，其次，利用大语言模型强大的生成能力对原问题的题干进行重写，进一步，利用问题词扩展库对其扩展，最后，通过对散文文章分块处理，建立基于全局上下文信息、历史信息的问题和文章句子的相关性判断模型，用于抽取答案候选句。通过在散文阅读理解数据集上进行实验，实验结果表明本文提出的方法提高了散文抽取答案候选句的准确率，为散文阅读理解的生成类问题的解答提供了技术支撑。

**关键词：** 文本抽取 ； 问题词扩展库 ； 散文体裁

# Sentiment classification method based on multitasking and multimodal interactive learning

**Yang Lei[1]，Suge Wang[1,2,*]，Shuqi Li[1]，Hao Wang[1]，**
1.School of Computer and Information Technology,
Shanxi University,Shanxi 030006
2.Key Laboratory Computational Intelligence and Chinese Information
Processing of Ministry of Education, Shanxi University,Shanxi 030006
wsg@sxu.edu.cn

## Abstract

In prose reading comprehension, on one hand, the questions are usually concise and abstract, making it difficult for machines to directly understand the meaning and requirements of the questions. On the other hand, prose passages are lengthy, and candidate answer sentences are dispersed across multiple paragraphs in the passage, posing significant challenges for sentence extraction tasks. Therefore, this paper proposes a question-enhanced approach for extracting candidate answer sentences from prose passages. Firstly, relevant words in the passage are extracted using a large model to build a question word extension library. Subsequently, the powerful generative capability of the large model is utilized to rewrite the original question, and then the question word extension library is employed for further expansion. Lastly, by segmenting the prose passage, a model is established to assess the relevance between questions and sentences based on global context information and historical information, facilitating the extraction of candidate answer sentences. Experimental results on a prose reading comprehension dataset demonstrate that the proposed method improves the accuracy of

第二十三届中国计算语言学大会论文集，第613页-第624页，太原，中国，2024年7月25日至28日。
卷1：主会论文
(c) 2024 中国中文信息学会计算语言学专业委员会

extracting candidate answer sentences from prose passages, thereby providing technical support for answering generative-type questions in prose reading comprehension tasks.

**Keywords:** text extraction , question expansion library , prose genre

## 1 引言

机器阅读理解是自然语言处理领域的一个重要研究方向，其目标是让机器理解文本中的语义并回答相应的问题。现有的机器阅读理解方法大多针对文本长度较短、问题形式相对固定的数据集。散文作为作者表达自身思想情感的载体，题材广泛、结构自由，具有深邃的意境和丰富的情感，使得针对散文的机器阅读理解任务更具挑战性。以表1的《遍野荆花》为例，从问题题干可以看出，题干内容精简，用词抽象，且与文章内容的字面呼应较少，若直接将问题与文章内容进行匹配时，导致机器难以理解问题中的抽象词，如“外部环境”，且问题与文章内容之间的关联困难，从而影响问题与文章内容的匹配，降低机器抽取答案候选句的准确率。另外，散文通常较长，我们对散文数据集SWQA统计发现，每篇散文平均包含1500字左右，且包含10个段落以上，而每个问题的答案候选句平均分布在3.5个段落以上，这将导致全面获取候选答案句较为困难。

针对阅读理解任务中问题题干有效信息较少的情况，现有研究(Yasunaga et al., 2021)通常引入知识图谱或外部知识对问题扩展，但对于散文体裁的阅读理解问题，题干用词抽象和多样，难以利用已有的知识或其他外部资源直接进行扩展。因此，本文利用大语言模型对原问题进行扩展和丰富语义。首先利用大语言模型抽取文章中与问题题干相关的词，构建问题词扩展库，再利用大语言模型强大的生成能力和丰富的词汇信息对原问题的题干进行重写。例如，表1中的问题，基于对文章内容的理解，大语言模型重写后的问题为“文章如何通过描绘恶劣的自然环境、贫瘠的土地资源和匮乏的农业资源等外部环境，深入展现农村的贫困？”。经过重写后，题干中融入了文章中的部分信息，使问题与文章的内容可以相呼应。随后调用问题词扩展库中具体的环境描绘词对“环境描写”做出解释。最终处理后的问题为“文章如何通过描绘恶劣的自然环境、贫瘠的土地资源和匮乏的农业资源等外部环境，深入展现农村的贫困？（水源缺乏，土壤贫瘠，怪石遍布，植被稀疏）”，由此可以看出通过题干重写和问题扩展，增强了问题与文章之间的联系。

对于散文长度较长、答案候选句分散的问题，预训练模型普遍存在输入长度受限的问题，如BERT模型最大输入长度为512个token，因此无法直接使用预训练模型应用于散文文本。研究者们一般采用滑动窗口(Zhang et al., 2023)、直接截断(Zhang et al., 2023)的方法处理长文本。而直接截断的方法势必会丢失一部分文章信息，影响模型的召回率。滑动窗口虽然分段将文本输入到模型中，但是窗口分割可能导致文本中重要的语境信息被切分，降低模型对整体文本的理解能力。因此，在一次不能将文章全部输入的情况下，我们参考文献(侯祺积, 2023)，将文本分为多个子块，若文章的长度超过每个子块的长度，将剩余部分划分到下一个子块中，以此类推，直至将长文本处理结束。面对答案候选句分散的情况，常见的多片段抽取模型(Deng et al., 2021; Yang et al., 2021)存在抽出的句子重复率较高、准确率较低的问题。因此，本文提出了问题与文章句子相关性判断模型，引入了历史上下文信息编码模块，在待抽取句子和已抽取句子间使用自注意力机制，降低了模型抽取答案候选句的冗余率。最后，将问题与分块后的散文输入到相关性判断模型中，实现散文答案候选句抽取。

本文的主要贡献如下：

（1）针对散文问题题干较为抽象的特点，设计了基于大语言模型的问题重写策略，同时利用大语言模型文本生成能力，构建了问题词扩展库，增加了问题的有效信息，丰富了问题的题干。

（2）针对候选句散落在散文的多个段落中，提出了问题与文本相关性判断模型，使模型可以捕获文本语句间的长距离依赖关系，提高了与问题内容相关的答案候选句的抽取性能。

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版
* 通讯作者 Corresponding Author.
**基金项目：**国家自然科学基金项目(62076158,62376143)

| | |
|---|---|
| 文章 | ……<br>……崮崖是个小山村，全村500多口人，只有不到300亩山岭地，<br>却有6000多亩山场。可这么多山场有什么用?<br>这是水源缺乏、土壤贫瘠、几乎连一棵大树都长不起来的山岭呀！……<br>……<br>……站在这里放眼四望，山野怪石遍布，植被稀疏。<br>多数植被是一种叫荆棵的低矮灌木，偶有几根针叶松。<br>刺槐之类的，也长得歪歪扭扭，一副苦大仇深的架势，他的心里更加迷茫了……<br>……<br>崮崖村生产的荆花蜜，色如纯净琥珀，入口留香绵长，投放市场后供不应求。<br>从此，这遍野荆花成为村里永不枯竭的财富之源……（有删改） |
| 问题 | 文章通过哪些外部环境描写展现了农村的贫困? |
| 候选句 | 崮崖是个小山村，全村500多口人，只有不到300亩山岭地，却有6000多亩山场。<br>可这么多山场有什么用？这是水源缺乏、土壤贫瘠、几乎连一棵大树都长不起来的山岭呀！<br>站在这里放眼四望，山野怪石遍布，植被稀疏。<br>多数植被是一种叫荆棵的低矮灌木，偶有几根针叶松。<br>刺槐之类的，也长得歪歪扭扭，一副苦大仇深的架势，他的心里更加迷茫了。 |

表 1: 散文《遍野荆花》节选

（3）在散文阅读理解数据集上，将本文方法与对比方法进行了实验，结果表明本文方法的性能优于其他对比方法。

## 2 相关工作

对于机器阅读理解任务，开放域问答是此类应用研究的重要组成部分，该任务是通过从大量文档中收集证据并回答相应的问题。针对散文阅读理解的答案候选句抽取，需要多片段文本抽取技术的支持，目前常用的机器阅读理解的方法是基于预训练模型方法以及结合外部知识的抽取技术。

在基于预训练模型的抽取式机器阅读理解方法中，Cui等人(2020)在RoBERTa的基础上，提出了MacBERT模型，该模型利用相似的单词做MASK掩码，缩小了预训练和微调阶段间的差距，且在多个中文NLP数据集上取得显著效果，因此，MacBERT可用于散文阅读理解的数据编码；Deng等人(2021)提出ReasonBERT，通过增强模型在长文本和多重上下文的推理能力，可为解决散文长文本的问题提供思路。Khattab和Zaharia (2020)提出了COLBERT模型，不同于常见的All-to-all交互方式，该模型具有延迟交互的能力，以提高问题和文本之间的交互效率。Yavuz等人(2022)提出了PATHFID模型，线性化支持段落的分层推理路径及关键句子，在提升模型性能的同时，使得到的问题候选答案句更具可解释性。Zhan等人(2021)采用对比学习的思想，提出了动态硬负采样的训练策略，在保持较好稳定性的同时，能够关注排名靠前的答案候选句。Gu等人(2022)提出了利用多步骤情节性马尔可夫决策过程抽取长文本的方法。侯祺积(2023)提出了Top-MRV模型，引入问题词扩展集来划分子句权重，提升了模型抽取答案候选句的准确率。

近年来，将外部知识引入MRC模型成为学者们的研究热点，形成了基于外部知识库的机器阅读理解（Knowledge-Based MachineReading Comprehension，KBMRC）。KBMRC与MRC的主要区别在于输入，MRC的输入是文本和问题序列，而KBMRC(Wang J et al., 2022)的输入除此之外还从知识库中提取出额外相关知识。Wang等人(2023)利用知识引导的提示作为干预，提升了模型在文本概念提取任务的准确率。

上述模型直接用于散文阅读理解任务，存在理解散文文章语义不全面，抽取出的句子准确率较低、冗余率较高等问题。而散文阅读理解中的问题与文章内容有密切的关系，若利用这些信息可以帮助模型实现答案候选句的抽取。例如，散文中对人物具体的描绘、环境的描写，这些信息可以用于丰富问题题干，提供额外的语义信息，以此提升散文答案候选句的抽取性能。


(c) 2024 中国中文信息学会计算语言学专业委员会

## 3 散文问题扩展和答案候选句抽取模型

针对散文答案候选句的抽取任务，受文献(侯祺积, 2023)构建多片段答案关键句抽取框架方法的启发，一方面，利用大语言模型和问题抽象词扩展库对问题进行重写和扩展，另一方面，将文章进行分块处理，并将分块后的文章和问题输入到相关性判断模型。在相关性判断模型中，全局上下文编码模块对输入数据进行编码，得到文章句子的全局向量表示。历史信息编码模块包含两个多头自注意力层，分别捕获待抽取句子间的交互信息和待抽取句子与已抽取句子间的交互信息。两层自注意力层的最终输出构成待抽取句子的历史信息表示。将每个待抽取句子的全局向量与历史信息向量进行连接，形成聚合向量。聚合向量经过映射后得到每个问题的答案候选句得分，选取得分最高的句子作为答案候选句。模型整体框架如图1所示。

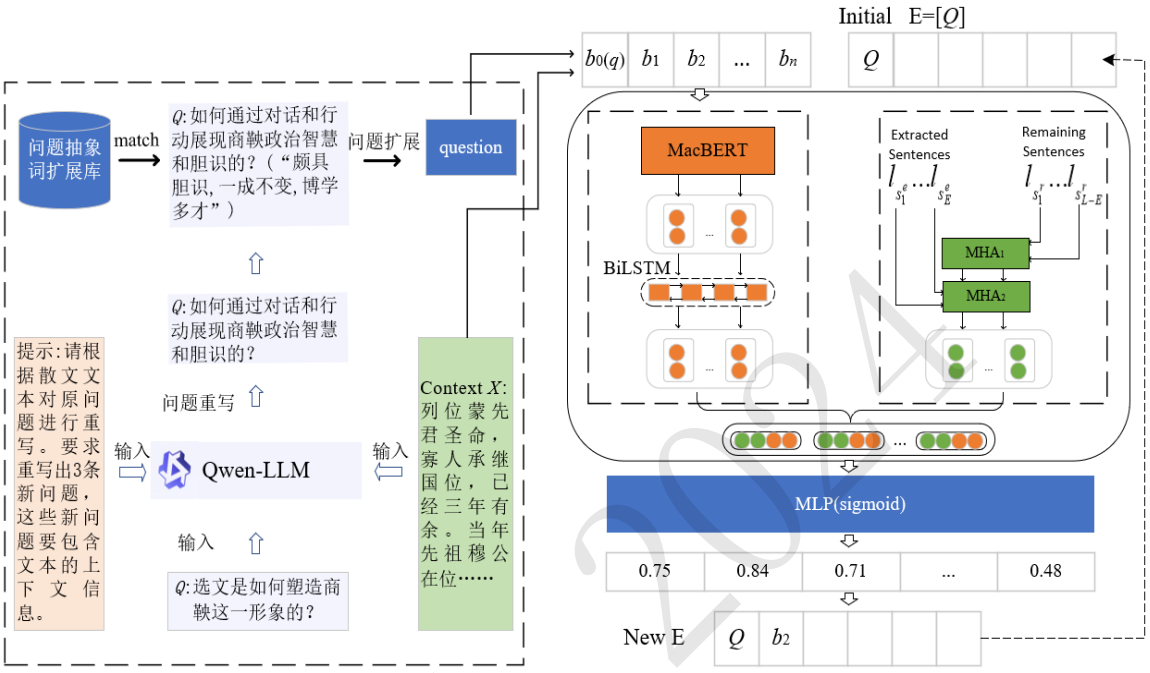



图 1: 散文问题扩展与答案候选句抽取模型框架图

### 3.1 基于大语言模型的问题扩展

由于散文阅读理解中的问题简洁、用词抽象，影响了答案候选句的抽取结果，为此，本文研究基于大语言模型的问题扩展。主要包括问题词扩展和问题重写。

#### 3.1.1 问题词扩展库的建立

通过分析数据集中的各类问题，我们发现在散文人物形象类和环境描写类问题中，经常出现一些抽象词，如“形象”和“环境描写”，模型对这类抽象词的理解较为困难，进而可能会影响问题与文章内容的相似度匹配，导致散文答案候选句抽取的质量下降。为此，可以根据问题题干中提及的具体人物，从文章中找到相关的人物形象描述，将这些人物描写词扩展到问题中。类似地，针对环境描写类问题，从文中找到相关的环境描写词，扩展到问题中，以具体化抽象词，使机器更好地理解抽象概念。

本文使用千问大语言模型构建问题扩展库，具体过程如下：

（1）人物形象和环境描写词汇抽取：利用大语言模型识别出文本中出现的人名实体，然后针对各人物抽取对应的人物形象词。类似地，抽取描写环境的词语。

（2）扩展词选取：在散文文章中，一方面，对一些人物的描述较为简略，导致可供抽取的形象词数量较少；而对主要人物往往赋予较多的描写，使得可抽取的形象词数量较多。因此，在文章中不同人物的描写词数量存在差异。为了获取问题的扩展词，本文制定以下策略：选择每篇散文中人物描写或环境描写的前4个词作为扩展词。若描述词数量少于4个，则会全部选取作为问题扩展词。

例如，在《商鞅变法》一文中，问题: “选文是如何塑造商鞅这一形象的？”，该文中并未直接提及“形象”这一词汇，题干中“形象”一词与全文内容匹配度较低，但文章中出现了对商鞅具体的人物描写，若将这些形象词扩展到问题中，对抽象词起到解释的作用。扩展后的问题将变为：“选文是如何塑造商鞅这一形象的？(“颇具胆识，一成不变，博学多才，贸然行事”)”。对于环境描写类问题，采用类似的扩展方法。最终，本文共建立了7140条问题扩展词，此外，需要注意的是本文的问题词扩展库是以文章为组别的。样例如表2所示。

| 人物对象/环境 | 人物描写/环境描写 |
|---|---|
| 商鞅 | 颇具胆识，一成不变，博学多才，贸然行事 |
| 秦孝公 | 奇耻大辱，不悦，殷纣夏桀，激动，易弦更张 |
| 公孙大人 | 妄议朝政，巧舌如簧，腐朽 |
| 公子虔 | 博学多才，不拘古法 |
| 祝欢 | 早有征兆，不敢妄言 |
| 环境 | 暴风雪，零下三十摄氏度，温度降低，狂风 |

表 2: 问题抽象词库示例

### 3.1.2 问题重写

通过对散文数据的分析，绝大部分散文阅读理解问题的题干较为简洁，例如：“选文是如何塑造商鞅这一形象的？”，问题中的关键信息为“商鞅”和“形象”，仅依靠这两个词语，模型无法有效学习到问题题干与文章内容的联系，影响了模型对答案候选句的判断。因此，针对题干中有效信息较少的问题，本文利用大语言模型强大的文本生成能力，构建基于大语言模型的问题重写过程，以增强并扩充题干中的上下文信息。具体过程如下：

（1）提示信息建立：依据文章内容对原问题进行重写，要求对原问题重写出3条问题，这些重写的问题要求包含文本的上下文信息。

（2）问题重写：将原问题、文章内容和提示信息作为输入，利用阿里云的通义千问大语言模型，生成与原问题语义相似的问题重写，这些问题重写可能在表达方式、结构或语气上与原始问题有所不同，但在保持原问题相同的语义和意图的基础上，增加了题干中的信息量，丰富了原问题，具体形式如下所示：

$$\{Q_{\text{rewrite}}\} = \text{Qwen}\left([Q; \text{Context}; \text{Prompt}]\right)$$

其中，$\{Q_{rewrite}\}$表示重写问题集合。

（3）问题筛选：在生成新问题后，进行问题筛选。将重写后的问题与文章内容、原问题分别做相似度计算，将二者计算后的结果求和排序，选取相似度得分最高的问题作为重写后的问题。

本文使用基于TFIDF(张兆滨等, 2020)的相似度计算方法，具体计算见公式(1)-(4)：

$$tf_{S,w_i} = \frac{C_{S,w_i}}{\sum_k c_{k,w_i}} \tag{1}$$

$$idf_{w_i} = \log(\frac{|D|}{|1 + \{S : w_i \in S\}|}) \tag{2}$$

$$tfidf_{S,w_i} = tf_{S,w_i} \cdot idf_{w_i} \tag{3}$$

其中，$C_{S,w_i}$是词$w_i$在句子$S$中出现的次数，$|D|$是句子总数，$|\{S : w_i \in S\}|$是包含词$w_i$的句子数。$tf_{S,w_i}$和$idf_{w_i}$分别为词$w_i$的词频和逆文档频率。

重写问题$Q$与文章句子$S_g$的相似度$\cos(Q, S_g)$，具体计算见公式(4)。


(c) 2024 中国中文信息学会计算语言学专业委员会

$$\cos(Q, S_g) = \frac{\sum_{i=1}^{d_w} (tfidf_{Q,w_i} \times tfidf_{S_g,w_i})}{\sqrt{\sum_{i=1}^{d_w} (tfidf_{Q,w_i})^2} \times \sqrt{\sum_{i=1}^{d_w} (tfidf_{S_g,w_i})^2}} \tag{4}$$

其中$tfidf_{Q,w_i}$表示词$w_i$在问题$Q$中的TFIDF值，$tfidf_{S_g,w_i}$表示词$w_i$在文章句子$S_g$中的TFIDF值，$d_w$表示词向量的维度。

类似地，可以定义重写问题$Q$与原问题$q$的相似度$\cos(Q, q)$。

对两类信息的相似度得分求和，得到每个重写问题最终的得分，具体计算见公式(5)。

$$grade(Q, S_g, q) = \cos(Q, S_g) + \cos(Q, q) \tag{5}$$

对于《商鞅变法》一文，原问题为“选文是如何塑造商鞅这一形象的？”，重写后的问题为“如何通过对话和行动展现商鞅的政治智慧和胆识？”。重写后的问题相比于原问题内容更具体详尽，同时引入了文中的部分相关情节，提供给模型更多的上下文信息。

### 3.2 答案候选句抽取模型构建

为了实现答案候选句的抽取，设计了相关性判断模型，该模型首先利用文本分块处理散文长文本，其次，利用全局上下文信息和历史信息编码模块对问题与句子间的关系做出判定，最后根据判定分数，选取答案候选句。

（1）分块处理

预训练语言模型MacBERT在处理输入序列时，通常有长度限制，为了最大程度地输入文本内容，对文本进行了分块处理。具体地，将文本分成$n$个子块，将问题$q$记作子块$b_0$，文本记为子块$b_1, ..., b_n$。每个子块的长度为$l = length/n$，其中，$length$为整个文本的长度。对于每个子句$s_k$，若长度超过了$l$，则将超出部分的子句划分到下一个子块$b_{i+1}$中。同时，下一个子句$s_{k+1}$将划分到$b_{i+2}$中，以此类推，直到子块的个数达到$n$，如图2所示。

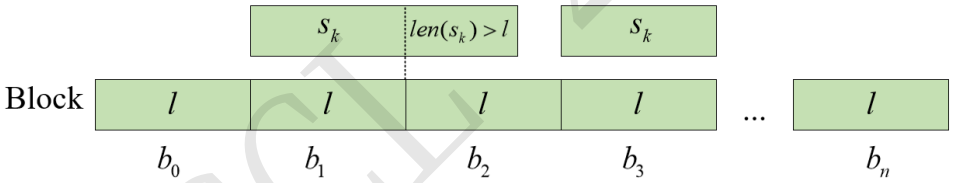



图 2: 子句超出固定长度$l$的情况

在数据处理中，对于文本的分块处理是关键的。考虑到问题句与散文文本之间的相关性可以为答案候选句抽取提供重要的信息。因此，在训练数据集中对答案候选句进行了标注，用于表示Block中文本与问题之间的相关性，具体形式如下所示：

$$R(Block) = [\underbrace{1, 1, ..., 1}_{b_0(q)}, \underbrace{0, 0, ..., 0}_{b_1}, \underbrace{1, 1, ..., 1}_{b_2}, ...] \in [0, 1]^{len(Block)}$$

其中，Block中的第一块$b_0$代表问题块，接着的$n$个子块则代表文本块。在$b_1$到$b_n$中，如果$b_i$是答案候选句，则标注为1,1,...,1；否则标注为0,0,...,0。

（2）相关性判断模型

相关性判断模型由全局上下文编码模块、历史信息编码模块和MLP三部分组成。

全局上下文编码模块：设初始文本序列为$X$，首先使用MacBERT(Cui et al., 2020)，将块中信息映射到一个高维向量空间中，获得包含文本信息的向量表示M，在此基础上，使用BiLSTM(Huang et al., 2015)建立各块间的前后语义关联，获得各块中句子的上下文信息表示$G_{S_i}$，具体表示见公式(6)-(8)：

$$X = (Q, b_1, b_2, ..., b_n) \tag{6}$$


(c) 2024 中国中文信息学会计算语言学专业委员会

$$M = \text{MacBERT}(X) \tag{7}$$

$$G_{s_i} = \text{BiLSTM}(M) = (l_{s_1}, ..., l_{s_n}) \tag{8}$$

历史信息编码模块：设已抽取句子序列为$E$，待抽取句子序列为$V$，首先在待抽取句子间执行多头自注意力，以使每个待抽取句子捕获其他剩余句子提供的上下文信息，记作$V_{MHA}$，其次在$V_{MHA}$和已抽取句子序列$E$之间执行多头自注意力，得到每个待抽取句子的历史信息表示$H_{s_i^r}$。连接待抽取句子的全局上下文表示$G_{s_i}$和历史信息表示$H_{s_i^r}$，得到拼接后的向量$H_i^r$，具体表示见公式(9)-(13)：

$$E = (l_{s_1^e}, ..., l_{s_E^e}) \tag{9}$$

$$V = (l_{s_1^r}, ..., l_{s_{n-E}^r}) \tag{10}$$

$$V_{MHA} = \text{MHA}_1(V, V) \tag{11}$$

$$H_{s_i^r} = \text{MHA}_2(V_{MHA}, E) \tag{12}$$

$$H_i^r = \text{concate}(G_{S_i}, H_{s_i^r}) \tag{13}$$

将拼接后的向量$H_i^r$输入到MLP中，再通过sigmoid函数得到问题和文章句子间的相关性得分$estimate(s_i)$，具体表示见公式(14)：

$$\text{estimate}(s_i) = \text{sigmoid}(W H_i^r + b) \tag{14}$$

其中,$W$为可训练参数，$b$为偏置，$estimate(s_i) \in [0, 1]$。

（3）获取答案候选句

根据判断模型计算出的问题和句子间的相关性得分$estimate(S)$，选择得分较高的前k个子句作为最终答案候选句。

### 3.3 模型优化

该模型的优化目标是使预测的分数$estimate(S)$与标准分数无限接近，本文采用了最小化交叉熵损失的方法，损失函数如式(15)所示。

$$\text{Loss}_{\text{estimate}} = \text{CrossEntropy}(\text{estimate}(S), R(\text{Block})) \tag{15}$$

其中,$R(Block)$表示问题与文本句子间的真实标签。

## 4 实验

### 4.1 数据集

本文采用散文阅读理解数据集SWQA，该数据集由两部分构成：它是由历年高考语文真题、模拟题、网络资料、线下纸质题集，数据集共3790篇，记作SWQA_1；科大讯飞提供的组卷网上的散文阅读理解数据集共11560篇，记为SWQA_2；将散文数据集的SWQA_1和SWQA_2进行随机排序合并，再采用json格式进行存储，合集记为SWQA。共录入数据15350份，将其按照8：1：1的比例随机划分为训练集、验证集和测试集，如表3可见。

由于散文篇幅较长，答案分散，题型类型多样。根据题型的特定关键词，将散文的阅读理解问题分为观点理解类、概括类、作用类、原因类、表现手法类、标题理解类、赏析类七类，如表4可见。

按照各类别对散文的篇幅平均长度、平均段落量、答案在段落中平均分布量进行统计，结果如表5可见。

CCL 2024

| 数据集 | 数据量 | 训练集 | 验证集 | 测试集 |
|---|---|---|---|---|
| SWQA_1 | 3790 | 3032 | 400 | 358 |
| SWQA_2 | 11560 | 9245 | 1140 | 1175 |
| SWQA | 15350 | 12277 | 1540 | 1533 |

表 3: 数据集规模介绍

| 问题类别 | 特定关键词 |
|---|---|
| 概括类 | 概括、特点、形象、特色等 |
| 观点理解类 | 观点、理解、启示等 |
| 作用类 | 作用、妙处、好处等 |
| 原因类 | 原因、为什么、解释等 |
| 表现手法 | 表现手法、手法、描写手法等 |
| 标题理解类 | 标题、题目等 |
| 赏析类 | 赏析、鉴赏等 |

表 4: 题型分类关键词示例

### 4.2 评价指标及参数设置

本文使用精确率P（Precision）、召回率R （Recall）和F1 值（F1 Score）作为模型的评价指标。实验采用Pytorch 1.12.1 框架，GPU 选用NVIDIA TITAN RTX。具体参数设置如表6所示。

### 4.3 实验设计

为了验证本文中的答案关键句抽取方法的有效性，我们选择了一些片段抽取任务中的模型作为对比实验，具体模型如下：

DRhard(Zhan et al., 2021)：基于RoBERT的编码层，分别将查询和文档编码为低维嵌入，并使用查询嵌入进行高效的相似性搜索进行在线排名，此外硬负采样通过最小化前K对误差，进一步改进了排名性能。

ColBERT(Khattab and Zaharia., 2020)：将查询和文档编码为BERT上下文的嵌入，通过计算它们之间的相似性来评估相关性。此外，ColBERT的后期交互机制支持直接从大型文档集合中进行端到端的检索，进一步提高了检索效率和召回率。

MonoQA(Kongyoung et al., 2022)：查询和文档经由RoBERT构成的ConvDR检索模型编码为向量形式，这些向量经过点积运算以后得到Top-K个文档，将Top-K个文档连同查询经由Reranker再做一次排序，得到最终要选择的文档。

Block-Skim(Guan et al., 2022)：对Transformer模型进行改进，引入了一个注意力权重热图，并对其进行了处理。利用卷积神经网络来处理这个热图的对角线部分，以确定文本是否需要被保留。我们设置了一些参数，learning rate：$3e^{-5}$，batch size：8，seed：42，训练周期：50，最大序列长度：512，步幅：12。

CogLTX(Ding et al., 2020)：结合BERT（BERTbase）与多层感知机制的循环迭代模型，参数：learning_rate：$1e^{-4}$，batch size：8，weight_decay：0，epochs：70，Attention heads：12。

Top-MRV(侯祺积, 2023)：以MacBERT为编码器，引入了问题词扩展用于划分子句的初

| 问题类别 | SWQA_1 | SWQA_2 | 字数/篇 | 段落/篇 | 答案分布段落/篇 |
|---|---|---|---|---|---|
| 概括类 | 739 | 2248 | 1518.62 | 11.75 | 3.87 |
| 观点理解类 | 527 | 2554 | 1517.90 | 11.12 | 3.76 |
| 作用类 | 723 | 2071 | 1577.44 | 10.57 | 4.69 |
| 原因类 | 534 | 1236 | 1419.15 | 10.34 | 3.58 |
| 表现手法 | 381 | 938 | 1483.73 | 10.54 | 3.63 |
| 标题理解类 | 468 | 1102 | 1545.11 | 11.66 | 4.46 |
| 赏析类 | 418 | 1411 | 1597.54 | 12.09 | 4.73 |

表 5: SWQA_1和SWQA_2题型分布比例


(c) 2024 中国中文信息学会计算语言学专业委员会

| 超参数名 | 参数值 |
|---|---|
| Dropout | 0.1 |
| Learning rate | 0.0001 |
| Batch size | 8 |
| Hidden size | 768 |
| Hidden layer | 12 |
| Max sequence length | 512 |
| Weight decay | 0.0001 |
| Attention heads | 12 |

表 6: 超参数设置

始权重，将初始权重与预测的子句权重动态调整，增强了模型抽取答案关键句的能力，基线模型。

### 4.4 实验结果及分析

#### 4.4.1 实验比较及分析

将本文的方法与第4.3节的六种方法进行对比实验，实验结果如表7所示。

| Model | P | R | F1 |
|---|---|---|---|
| ColBERT | 35.17 | 39.40 | 30.27 |
| DRhard | 39.10 | 39.23 | 35.13 |
| MonoQA | 49.91 | 37.06 | 42.53 |
| Block-Skim | 58.79 | 49.84 | 50.04 |
| CogLTX | 59.22 | 56.71 | 55.63 |
| Top-MRV | 70.21 | 64.39 | 65.22 |
| OURS | 73.54 | 64.46 | 67.00 |

表 7: 本文方法与六种对比方法的实验结果

通过表7可以看出：

（1）本文提出的模型在精确率P、召回率R和F1值评价指标中均优于其他六种模型，验证了本文方法在散文中抽取答案候选句的有效性。

（2）本文提出的模型相比于检索模型ColBERT和DRhard，在F1值上分别提高了36.73和31.87。对于ColBERT模型，针对散文文本较长时未考虑文本分块，缺乏长距离语义交互，导致模型表现不佳。另外，虽然DRhard模型是对句子进行处理，在F1值上高出ColBERT模型3.93，但是所有文章的句子全都打乱混在一起，增加了搜索的难度，模型效果也较差。

（3）本文提出的模型相比于MonoQA模型在F1值上提高了24.77，主要原因是MonoQA模型是用于阅读理解生成，无法发挥多任务学习的效果。

（4）相较于Block-Skim和CogLTX模型，本文模型在F1值上分别提升了16.96和11.37。两者都对长序列进行了分块处理，其中Block-Skim使用了注意力权重热图的对角线部分判断答案是否保留，这种模型虽然提高了模型的运行速度，但是对于联系上下文更为密切的散文阅读理解任务而言，其表现并不理想。而CogLTX利用判断模型对文章选择性截取来实现缩短长文的效果，但是经过处理后的文章丢失了语句间的关联性，给阅读理解任务带来了语义理解上的困难。

（5）相比于Top-MRV模型略有提升，主要原因是本文模型利用大语言模型对原问题进行了重写，并引入了问题抽象词扩展库，具体化了问题题干的抽象词，丰富了原问题的有效信息量，此外，连同相关性判断模型的引入，共同提升了答案候选句抽取的准确率。

第二十三届中国计算语言学大会论文集，第613页-第624页，太原，中国，2024年7月25日至28日。
卷1：主会论文
(c) 2024 中国中文信息学会计算语言学专业委员会

#### 4.4.2 消融实验

为了验证本文模型在各个部分的性能，本文设计了消融实验。

(1)-Block：将本文模型去除分块模块；

(2)-Rewriting Question：将本文模型去除问题重写模块；

(3)-Abstract feature：将本文模型去除问题抽象词扩展库；

(4)-Judge：将本文模型去除相关性判断模型；

(5)-History Encoder：将本文模型去除历史信息编码模块；

消融实验结果如表8所示。

| Model | P | R | F1 |
|---|---|---|---|
| ours | 73.54 | 64.46 | 67.00 |
| -Block | 65.23 | 59.24 | 62.09 |
| -Rewriting Question | 72.63 | 64.02 | 66.20 |
| -Abstract feature | 72.46 | 63.40 | 65.93 |
| -Judge | 45.62 | 70.22 | 55.25 |
| -History Encoder | 72.56 | 63.70 | 66.08 |

表 8: 消融实验结果

由表8可以看出，模型去除任意子模块，其综合性能都有一定的下降，模型去除各个模块后的表现如下：

（1）-Block相比于本文方法的P、R和F1值分别下降8.31、5.22、4.91，可以看出P值下降较大。主要原因是文本分块保留了散文的上下文，去除该模块以后丢失了大量文本信息，造成模型P值大幅下降。

（2）-Rewriting Question，相比于本文方法的P、R和F1值分别下降0.91、0.44、0.8，这是因为问题题干中的有效信息量减少，模型无法有效学习到问题题干和文本之间的关系，造成模型的性能下降。

（3）-Abstract feature，相比于本文方法的P、R和F1值分别下降1.08、1.06、1.07，是由于本文模型中的问题抽象词扩展库可以帮助模型解决问题题干中难以理解的语义部分，提升了模型的语义理解能力。

（4）-Judge，相比于本文方法的P和F1值分别下降27.92和11.75，召回率达到最高，抽取出的句子的准确率较低且冗余率较高，因此，不能直接用于实际阅读理解答题中，应融合其他模块对文本进一步筛选。

（5）-History Encoder，相比于本文方法的P、R和F1值分别下降0.98、0.76、0.92，这是因为提取历史编码器可以从待抽取句子的角度对已抽取句子考虑，若去掉该模块，降低了抽取的准确性。

综合上述分析可知，将问题重写模块、问题抽象词扩展库模块、文本分块和判断模型四个部分融合起来，将有助于提高散文答案候选句的抽取能力。

#### 4.4.3 实例分析

为了直观地了解本文方法抽取答案候选句提升的效果，利用本文模型与对比方法中性能最优的方法Top-MRV的获得实验结果进行比较分析。选取阅读理解问题为：“文章通过哪些外部环境描写展现了农村的贫困？”。该问题的标注答案的候选句如下：(1)“崮崖是个小山村，全村500多口人，只有不到300亩山岭地，却有6000多亩山场。”；(2)“可这么多山场有什么用？这是水源缺乏、土壤贫瘠、几乎连一棵大树都长不起来的山岭呀！”；(3)“站在这里放眼四望，山野怪石遍布，植被稀疏。多数植被是一种叫荆棵的低矮灌木，偶有几根针叶松。”；针对上述例子，Top-MRV模型得到的正确答案候选句为(1)和(2)。本文的方法首先会对原问题进行重写和扩展，得到扩展后的问题为“文章如何通过描绘恶劣的自然环境、贫瘠的土地资源和匮乏的农业资源等外部环境，深入展现农村的贫困？（水源缺乏，土壤贫瘠，怪石遍布，植被稀疏）”。根据问题扩展后的结果，抽取的答案候选句为(1)、(2)和(3)。由此可见，通过问题扩展，本文方法可以抽取出Top-MRV模型未能抽取出的答案句(3)，说明经过问题扩展后，可以找出外部环


(c) 2024 中国中文信息学会计算语言学专业委员会

境描写对贫穷的隐式刻画，而没有经过扩展的问题，仅能依据问题中的显示描写如“农村”，对文章句子进行抽取。

## 5 结论

针对散文阅读理解答案候选句抽取问题，提出了基于问题扩展的散文答案候选句抽取方法。该方法融合了更多文本知识，利用大语言模型实现了问题重写，建立了抽象词扩展库。针对文本较长的特性，对散文文本进行了分块处理，最终构建了候选句识别模型，通过对比实验，结果表明该方法优于目前先进方法，通过消融实验，验证了问题重写、问题抽象词扩展库和判断模型，可以更加针对性地增强模型抽取散文答案候选句的能力，为散文阅读理解的生成类问答提供技术支持。

## 参考文献

Yiming Cui, Wanxiang Che, Ting Liu, et al. Revisiting pre-trained models for Chinese natural language processing[C]. In *Findings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020: 657–668.

Ming Ding, Chengyue Zhou, Haoyu Yang, et al. Cogltx: Applying BERT to long texts[J]. *Advances in Neural Information Processing Systems*, 2020, 33: 12792-12804.

Xiang Deng, Yu Su, Alyssa Lees, et al. ReasonBERT: Pretrained to reason with distant supervision[C]. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021: 6112–6127.

Ning Gu, Elliot Ash, Rüdiger Hahnloser. MemSum: Extractive summarization of long documents using multi-step episodic Markov decision processes[C]. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022: 6507-6522.

Yiyang Guan, Zhiyuan Li, Zicheng Lin, et al. Block-Skim: Efficient question answering for transformer[C]. In *Proceedings of the Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence*, 2022, 36(10): 10710-10719.

Zhiheng Huang, Wei Xu, Kai Yu. Bidirectional LSTM-CRF models for sequence tagging[J]. *arXiv preprint arXiv:1508.01991*, 2015.

Wang J, Wang C, Qiu M, et al. KECP: Knowledge enhanced contrastive prompting for few-shot extractive question answering[C]. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022: 3152-3163.

Omar Khattab, Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over BERT[C]. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, 2020: 39-48.

Sangha Kongyoung, Craig Macdonald, Iadh Ounis. MonoQA: Multi-task learning of reranking and answer extraction for open-retrieval conversational question answering[C]. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022: 7207-7218.

Yang H, Sui D, Chen Y, et al. 2021. Document-level event extraction via parallel prediction networks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 6298-6308.

Yuan S, Yang D, Liu J, et al. 2023. Causality-aware Concept Extraction based on Knowledge-guided Prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics* , pages 9255-9272.

Masashi Yasunaga, Hongyu Ren, Antoine Bosselut, et al. QA-GNN: Reasoning with language models and knowledge graphs for question answering[C]. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021: 535-546.

Serkan Yavuz, Kazuma Hashimoto, Yanan Zhou, et al. Modeling multi-hop question answering as single sequence prediction[C]. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022: 974-990.


(c) 2024 中国中文信息学会计算语言学专业委员会

Jiaxin Zhan, Jiaxin Mao, Yiqun Liu, et al. Optimizing dense retrieval model training with hard negatives[C]. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021: 1503-1512.

Hao Zhang, Haoyu Song, Shuang Li, et al. A survey of controllable text generation using transformer-based pre-trained language models[J]. *ACM Computing Surveys*, 2023, 56(3): 1-37.

侯祺积. 散文阅读理解简答题的解答方法研究[D]. 太原: 山西大学硕士学位论文, 2023.

张兆滨, 王素格, 陈鑫, 赵琳玲, 王典. 阅读理解中观点类问题的扩展研究[J]. 中文信息学报, 2020, 34(6): 89-96, 105.

CCL 2024

# 基于预训练模型与序列建模的音素分割方法

**杨尚龙[1,2]，余正涛[*1,2]，王文君[1,2]，董凌[1,2]，高盛祥[1,2]**
1.昆明理工大学，信息工程与自动化学院，昆明，650500
2.昆明理工大学，云南省人工智能重点实验室，昆明，650500
1054179997@qq.com, ztyu@hotmail.com, 175360805@qq.com,
46761956@qq.com, gaoshengxiang.yn@foxmail.com

## 摘要

音素分割作为语音处理领域内的一个重要任务，对于关键词识别、自动语音识别等应用具有至关重要的意义。传统方法往往独立预测每一帧音频是否为音素边界，忽视了音素边界与整个音频序列以及相邻帧之间的内在联系，从而影响了分割的准确性和连贯性。本文提出一种基于预训练模型与序列建模的音素分割方法，在HuBERT模型提取声学特征的基础上，结合BiLSTM捕捉长期依赖，再用CRF优化序列，提升了音素边界检测的性能。在TIMIT和Buckeye数据集上的实验表明，本文方法优于现有技术，证明了序列建模在音素分割任务中的有效性。

**关键词：** 音素分割；自监督预训练；条件随机场；序列建模

# Phoneme Segmentation Based on Pre-trained Model and Sequence Modeling

**Shanglong Yang[1,2], Zhengtao Yu[*1,2], Wenjun Wang[1,2], Ling Dong[1,2], Shengxiang Gao[1,2]**
1. Faculty of Information Engineering and Automation,
Kunming University of Science and Technology, Kunming 650500, China
2. Yunnan Key Laboratory of Artificial Intelligence,
Kunming University of Science and Technology, Kunming 650500, China
1054179997@qq.com, ztyu@hotmail.com, 175360805@qq.com,
46761956@qq.com, gaoshengxiang.yn@foxmail.com

## Abstract

Phoneme segmentation is a crucial task in the field of speech processing, essential for applications such as keyword spotting and automatic speech recognition. Traditional methods typically predict phoneme boundaries independently for each audio frame, overlooking the intrinsic relationships between the phoneme boundaries, the entire audio sequence, and adjacent frames. This oversight can compromise the accuracy and coherence of segmentation. We proposes a phoneme segmentation method that leverages pre-trained models and sequence modeling. Based on acoustic features extracted by the HuBERT model, our method utilizes a BiLSTM to capture long-term dependencies and a CRF to optimize the sequence, thereby enhancing the accuracy and coherence of phoneme boundary detection. Experiments conducted on the TIMIT and Buckeye datasets demonstrate that our method outperforms existing techniques, validating the effectiveness of sequence modeling in phoneme segmentation tasks.

**Keywords:** Phoneme Segmentation , Self-supervised Pre-training , Conditional Random Fields , Sequence Modeling

*余正涛（通信作者）：ztyu@hotmail.com
**基金项目**：国家自然科学基金（U21B2027, 62376111, U23A20388, ）；云南省重点研发计划（202303AP140008, 202302AD080003, 202401BC070021）；云南省科技人才与平台计划（202105AC160018）
©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

## 1 引言

音素分割，亦称音素边界检测，是在语音信号中确定离散音素单元间时间边界的关键任务，该任务在关键词识别(Keshet et al., 2009)、语音识别(Rybach et al., 2009)、语音合成(Lee et al., 2021)等语音处理领域发挥着至关重要的作用，为这些应用提供了基础的信息框架。

音素分割可以根据是否使用监督信息，分为有监督和无监督两大类。在无监督场景中，音素边界的推断仅依赖于语音信号本身(Kreuk et al., 2020; Cuervo et al., 2022)。而有监督学习则利用已知的音素边界位置作为训练标签，进一步可分为文本相关监督和文本无关监督。文本相关监督需要额外的音素标签作为输入(McAuliffe et al., 2017; Lin and Wang, 2022)，并要求推断的边界位置与输入标签严格对齐，但这种强制对齐的方法在应用时也需要音素标签，限制了其在新语种上的适用性。

本文聚焦文本无关的监督学习，即仅使用音素边界作为监督信息，不涉及音素标签。以往的研究主要将此任务视为二分类问题，将音频帧分类为边界或非边界。这些传统方法通常条件独立地对每一帧进行预测(Franke et al., 2016; Zhu et al., 2022; Kim and Choi, 2023; Strgar and Harwath, 2023)，忽略了音频信号中帧与帧之间的时间依赖性。鉴于音频信号的连续特性，一帧音频是否为边界并非孤立事件，而是与其周围帧以及整个音频序列紧密相关，可见现有方法存在明显的局限性。

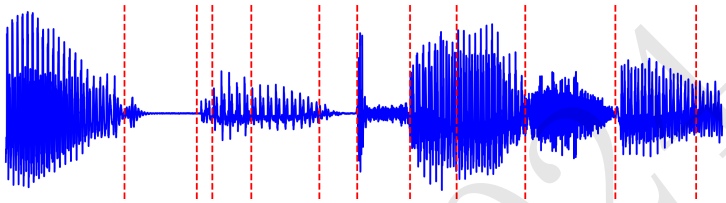


图1. 语音信号中的音素边界

受到序列标注任务的启发(Huang et al., 2015; Dalai et al., 2023; Zhou et al., 2019)，本文采用序列建模策略，提出了一种简单有效的音素分割方法。具体而言，首先使用预训练的HuBERT模型(Hsu et al., 2021)提取每一帧的音频表征，以利用无监督预训练过程中获得的丰富语音先验知识。然后，通过双向长短期记忆（BiLSTM）层(Graves and Graves, 2012)捕获音频序列中的长期依赖关系，进一步通过线性层将上下文表征转化为发射得分（emission scores），并利用条件随机场（CRF）层(Lafferty et al., 2001)对整个音频序列进行全局优化，以实现更加精确和连贯的音素边界判定。

本文主要贡献如下：（1）提出一种基于语音预训练模型与序列建模的音素分割新方法，强调了从整体序列角度进行建模的重要性；（2）在TIMIT(Garofolo, 1993)和Buckeye(Pitt et al., 2005)数据集上的实验证明，在不依赖额外音素标签信息的条件下，本文所提方法优于现有的音素分割方法；（3）本文方法具有良好的跨语言迁移能力，在训练未见的语言上直接推理也能取得较好的性能。

## 2 相关工作

### 2.1 音素分割

音素边界检测在多种实验设置下有着广泛的研究，具体可分为无监督、文本相关监督以及文本无关监督方法。

在无监督设置中，研究主要依赖于语音信号本身，不利用任何额外的音素或边界标签。传统方法侧重于利用信号处理技术，通过分析信号的频谱变化来探测音素边界(Hoang and Wang, 2015)。(Michel et al., 2016)利用循环神经网络（RNN）来预测下一帧音频，通过峰值检测算法发现可能的音素转变区域。(Kreuk et al., 2020)通过噪声对比估计技术来区分相邻帧和随机干扰帧，并应用峰值检测算法来定位音素边界。此外，对比预测编码（CPC）及其变体(Chorowski et al., 2021; Cuervo et al., 2022)在预测潜在向量时，尝试对未来多个潜在向量进行匹配，并使用对比损失进行模型训练，为无监督音素分割提供了新的思路。

在监督设置下，文本相关监督方法此前得到广泛研究。这类方法一般采用隐马尔可夫模型（HMM）或利用手工设计的特征进行结构化预测(Keshet et al., 2005; McAuliffe et al., 2017)。

CCL 2024

虽然文本相关方法在一定程度上可以提升音素边界的检测准确性，但其性能在很大程度上受限于对音素信息的需求，在没有可靠音素标签的新语言资源中存在较大的局限。

文本无关监督则为本文所关注的重点。在该设置下，只需要音素边界作为监督信息而无需提供具体音素标签。(Franke et al., 2016)使用LSTM来捕捉音频的上下文依赖关系，并通过交叉熵损失进行优化。(Kreuk et al., 2020)采用可学习的分段特征识别可候选的音素边界，通过动态规划算法来最小化结构化损失，还证明了来自音素标签的额外监督信号可以改善边界检测。(Zhu et al., 2022)提出了一种半监督方法，通过对比学习以及前向和损失直接学习语音到音素的对齐。(Kim and Choi, 2023)采用自回归架构来利用之前的边界预测信息，通过提供有关先前帧是否被分类为边界的附加信息来防止不必要的边界重复。(Strgar and Harwath, 2023)将迁移学习应用于音素分割任务，证明了自监督预训练模型的表征能力在音素分割任务上的有效性。

### 2.2 自监督预训练

自监督学习为一种无需标注数据的语音表示学习方法，通过自监督预训练从大量语音数据中学习有用信息，能够适应各类下游任务。自监督预训练模型近年来在语音领域取得显著成功，通过微调或作为特征提取器的方式在各项下游任务上取得先进的性能。

目前最广泛使用的预训练模型为wav2vec 2.0(Baevski et al., 2020)与HuBERT(Hsu et al., 2021)。wav2vec 2.0结合对比学习方法和掩码技术，旨在最大化上下文表征和本地化表征之间的相似性。在生成局部表征时，模型通过随机掩码学习使用未被掩盖的表征预测被掩盖部分，这样的设计使模型更关注输入波形的上下文信息，而不仅依赖局部表征，从而极大提高了泛化能力。HuBERT训练过程类似于声学模型和语言模型的联合学习：首先，模型学习从未被掩码的时间步提取连续潜在表征，这些表征被映射到离散单元，类似于传统的基于帧的声学建模。其次，HuBERT学习捕捉长期的时序依赖关系，以对掩码时间步做出准确预测。近期的研究证明，预训练模型编码的信息不仅包括传统方法所提取的帧级特征，还涵盖音素级、词级，甚至是语义层级的特征(Pasad et al., 2023)，对各种语音任务的研究具有深远的意义。

## 3 方法

本文所提出的模型架构包括HuBERT层，BiLSTM层和CRF层。在训练阶段，音频首先通过预训练的HuBERT模型，以获得声学表征序列。随后，BiLSTM层被用于进一步捕获音频序列中的长期依赖关系。其输出的上下文表征再通过线性层处理，输出每一帧对应各类标签的发射得分，这些得分表示了当前帧处于边界或非边界状态的概率，并用于决定最终的分类标签。CRF层则以全局方式对音频序列进行建模，综合利用每一帧的发射得分和标签间的转移概率，以模拟所有可能的标签序列，整个训练过程对HuBERT进行微调。在解码阶段，采用维特比算法来解码，得到最优的标签序列。

BiLSTM能够在每个时间点捕获过去和未来的信息，从而进一步增强预训练模型提取的特征，为每个时间步生成一个更全面的上下文依赖表征，在特征层面可以为CRF层提供更准确的发射得分；CRF层计算相邻标签转换的概率（例如从“边界”到“非边界”），并结合BiLSTM层输出的发射得分来最大化整体标签序列的概率。二者的结合能够利用BiLSTM捕获长距离依赖的能力和CRF全局序列优化的能力，这样我们的模型不仅能感知单帧的局部特征，还能理解整个音频序列的时序依赖性，从而提升音素分割的准确性和一致性。完整的模型架构如图2所示。

### 3.1 声学特征提取

在音素分割任务中，输入为原始音频序列$\bar{x} = (x_1, ..., x_T)$，序列的长度$T$与音频的总时长相关，原始序列通过特征提取转换为$\bar{z} = (z_1, ..., z_K)$这样的声学特征向量序列。每个输入语音序列都对应一个时间戳标签序列$\bar{y} = (y_1, ..., y_K)$，其中时间戳标签序列长度与声学特征向量序列长度$K$取决于预处理过程与特征提取方法，每个$y_i$是一个二元分类标签，用以标识音素边界或非边界。

传统的特征提取方法，如梅尔频率倒谱系数（MFCC），在捕获丰富的语音表征方面存在局限，特别是在提取音素级特征时。近年来，自监督学习在语音处理领域取得显著成果，在多个下游任务中展现了巨大潜力，通过对预训练模型进行微调或应用迁移学习至特定任务，已在众多语音处理任务上实现了前所未有的性能。


(c) 2024 中国中文信息学会计算语言学专业委员会

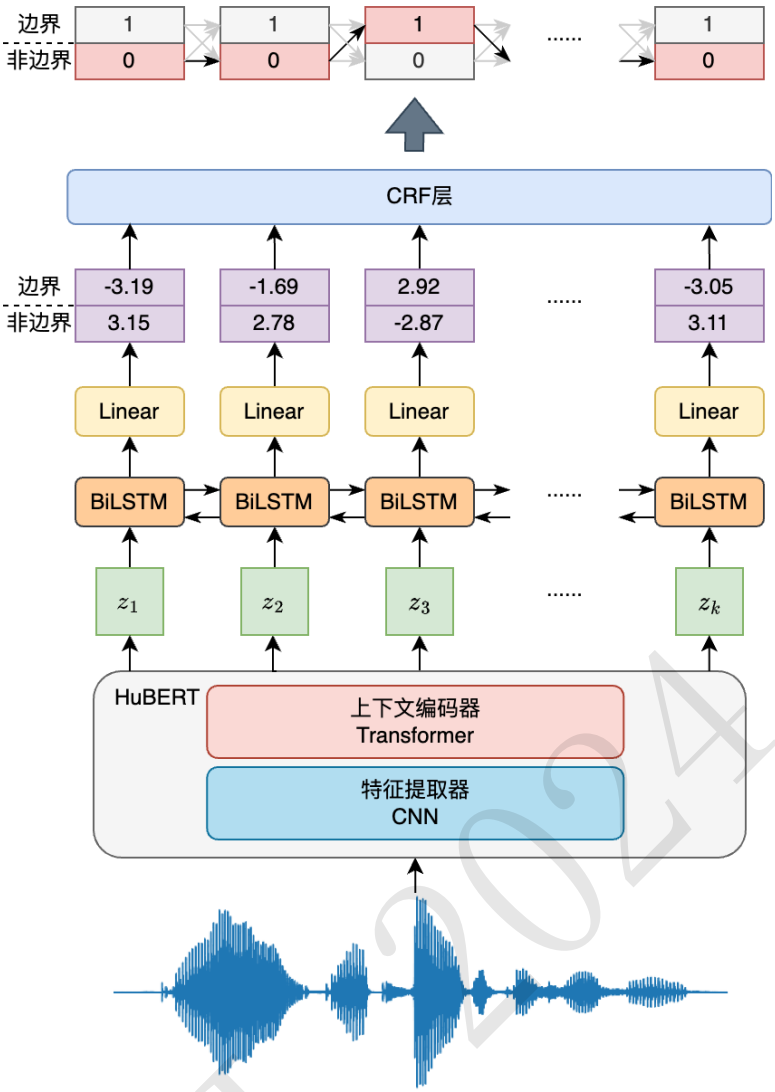



图2. 基于预训练模型与序列建模的音素分割架构

本文选择使用预训练的HuBERT模型作为声学特征的编码器。特征提取函数可以表示为$f \circ g$。其中$f$是一个卷积特征提取器，负责处理原始波形输入，并输出潜在的语音表示时间序列，这一序列在HuBERT的端到端训练过程中被学习；$g$是一个上下文编码器，它通过学习到的注意力掩码来从每个语音表征合成上下文感知的表征。函数$f$和$g$由HuBERT初始化，并在训练过程中接受梯度更新，从而优化特征表示，特征表示过程如式（1）所示。

$$\bar{z} = f \circ g(\bar{x}) \tag{1}$$

### 3.2 上下文信息建模

为了获得更丰富的上下文表征，本文采用双向长短期记忆网络（BiLSTM）进一步处理语音表征序列。长短期记忆网络（LSTM）是循环神经网络（RNN）的一种改良形式，设计用来学习序列数据中的长期依赖关系，有效地解决了梯度消失和梯度爆炸的问题。与RNN的主要区别在于，LSTM的隐藏层更新机制由专门设计的记忆单元替代，这些记忆单元负责控制信息的遗忘比例并将其传递到下一时间步。

在音素分割任务中，我们认为当前帧的预测不仅受到前序帧的影响，同样也与后续帧的信息密切相关。因此，我们引入BiLSTM来分别处理向前和向后输入序列，构建两种独立的隐藏状态，从而捕捉过去和未来的上下文信息。具体而言，可以将单向LSTM的隐藏状态更新过程表示为$\overrightarrow{h_t} = lstm(\overrightarrow{h_{t-1}}, e_t)$，其中$h_t$表示特定时间点的隐藏状态，$e_t$为当前时间步的输入。对于BiLSTM来说，我们得到两个方向的上下文表示，分别为：


(c) 2024 中国中文信息学会计算语言学专业委员会

$$\overrightarrow{h_t} = lstm(\overrightarrow{h_{t-1}}, e_t) \tag{2}$$

$$\overleftarrow{h_t} = lstm(\overleftarrow{h_{t-1}}, e_t) \tag{3}$$

其中$\overrightarrow{h}_t$和$\overleftarrow{h}_t$代表了从左到右（前向）和从右到左（后向）学习到的上下文信息。通过融合这两种方向的上下文表示，我们能够为每一帧音频构建更为全面的上下文表征。

$$h_t = \overrightarrow{h_t} + \overleftarrow{h_t} \tag{4}$$

### 3.3 标签序列解码

在音素分割任务中，利用相邻标签之间的依赖性或相关性对于减少独立预测每一帧音频可能产生的错误至关重要。例如，音素边界之后不太可能紧接着又出现一个边界，因此，利用相邻标签的相关性解码给定输入序列，可以确保生成的标签序列准确且合理。条件随机场（CRF）是一种在序列标注任务中广泛应用的方法，能够建模观测序列与状态序列之间的条件概率，在音素分割任务中，CRF在每个时间点的预测不仅考虑当前观测到的时间步，同时根据前后状态的依赖关系来优化整个序列的预测决策，它通过捕获序列中标签间的相互关系，优于单纯使用softmax函数的方法。我们将CRF用作标签序列解码器，以捕捉标签间的关联并对整个标签序列进行联合建模，而非单独预测每个标签。对于标签序列$\bar{y}$，我们定义第$i$个标签的得分为以下形式：

$$S_{y_i} = A_{y_{i-1},y_i} + P_{y_i} \tag{5}$$

这里的$A$是一个转移矩阵，其中$A_{i,j}$表示从标签$i$到标签$j$的转移得分，$P$是发射矩阵，由BiLSTM层的发射得分决定，因此，任何一个可能的标签序列的输出概率可以表示为：

$$p(Y_i \mid X) = \frac{exp(S(X, Y_i))}{\sum_{j=1}^{K} exp(S(X, Y_j))} \tag{6}$$

我们通过最大化对数似然函数来训练CRF，其损失函数的计算如下：

$$L_{CRF} = -\sum_{i=1}^{K} exp(S(y_i|x_i)) \tag{7}$$

通过这种方式，模型不仅能够有效地学习标签间的相互依赖性，还能够确保预测出的标签序列在整体上更加准确和连贯。

## 4 实验设置与结果分析

### 4.1 实验设置

#### 4.1.1 数据集

本文使用TIMIT和Buckeye语音语料库对模型进行训练和评估。对于TIMIT数据集，我们按照标准流程将数据划分为训练集和测试集，并仿照先前的研究(Strgar and Harwath, 2023)，使用训练数据的10%作为验证集。在处理Buckeye数据集时，我们根据之前的工作设定了训练集、验证集和测试集的比例分别为80%、10%和10%(Kreuk et al., 2020; Strgar and Harwath, 2023)。此外，为了提高实验的准确性和可靠性，我们将较长的录音文件在非语音噪声和静音期间切割为较短的连续语音片段，确保每个片段的起始和结束部分的非语音时间不超过20毫秒。这样的处理旨在优化数据输入，从而为音素分割任务提供更加稳定和准确的实验条件。

我们还在老挝语上进行了跨语言实验，在不进行额外训练的前提下直接使用老挝语测试集测试模型性能。测试集由老挝本土人员人工标注，为了保持实验的一致性，我们将数据标签转化为TIMIT和Buckeye相同的格式，所使用的公共数据集和老挝语测试集的具体信息如表1所示。

CCL 2024

| | TIMIT | Buckeye | 老挝语测试集 |
|---|---|---|---|
| 时长（小时） | 3.45 | 7.69 | 1.89 |
| 句子数（条） | 4009 | 10264 | 2050 |

表1. 数据集信息

#### 4.1.2 数据预处理

为了让音素标签与经过HuBERT预处理的音频数据之间保持时间上的一致性，我们对数据集的标签格式进行了必要的转换。原始数据集以采样点为单位，记录了每个音素的开始和结束时间，其中采样率为16kHz。目标是将这些时间戳标签转换成二进制的音素边界标签，生成的标签以20ms为单位，标识该帧是否为音素边界（边界标记为1，非边界标记为0）。

为实现这一转换，我们采取了以下步骤。首先，模拟HuBERT处理流程中的7层卷积操作，确定卷积后音频的长度，并计算原始音频长度与卷积处理后音频长度之间的比例因子。然后，利用这个比例因子，将原始音素时间标注转换为与处理后音频长度相匹配的新时间帧格式。通过这种方法，我们确保了音素标签与HuBERT处理后的音频数据在时间上的对齐，为后续的训练和评价提供了准确的标签信息。

#### 4.1.3 模型配置与评价指标

本文实验构建在fairseq(Ott et al., 2019)框架上，选择Librispeech(Panayotov et al., 2015)960小时数据集预训练的 HuBERT[1]作为基础检查点进行加载。在模型配置方面，我们设置BiLSTM的隐藏层维度为768，层数为2，对HuBERT层的微调学习率定为1e-4，而CRF层的学习率则设为1e-2。优化过程采用Adam优化器进行，其中$\beta_1$为0.9，$\beta_2$为0.999。所有的训练过程均在1张NVIDIA GeForce RTX 3090 GPU上进行，以16为批量大小，并设置默认训练周期为30。

评价指标方面，本文实验延续先前研究的做法，采用准确率、召回率、F1值以及R-value作为主要评价标准(Räsänen et al., 2009)，设置20ms为误差窗口。为了限制在一个误差窗口内，多个预测边界与单一真实边界匹配并重复计算贡献的情况，我们还采纳了(Strgar and Harwath, 2023)提出的更为严格的评价指标计算方法，即在多个预测边界与真实边界匹配的场景中，仅能有一个预测被视为正确并计数。考虑到R-value能够平衡召回率与过度分割的关系，其被认为是评估分割任务最为科学的指标。在模型训练期间，每处理50个批次的数据则在验证集上计算评价指标，根据严格指标下的最佳R-value保存最佳模型检查点，并用于最终测试。如果连续50次在验证集上的评估未能更新最佳模型将提前终止训练。

### 4.2 实验结果

#### 4.2.1 与现有音素分割方法的对比实验

为了验证本文方法的有效性，我们在TIMIT和Buckeye数据集上与现有的音素分割方法进行了对比。选取了以下几种方法作为基线模型，具体介绍如下：

（1）segfeat：利用可学习的分段特征来识别潜在的音素边界，并通过动态规划算法最小化结构化损失，结果取自(Kreuk et al., 2020)。

（2）superseg：采用自回归架构，预测时考虑前一帧的边界预测信息，并利用先前帧是否被分类为边界的附加信息来限制边界的重复预测，结果取自(Kim and Choi, 2023)。

（3）wav2vec2.0 finetune：通过二元交叉熵损失微调预训练的wav2vec 2.0进行音素分割，目前为这两个数据集上表现最佳的模型，结果取自(Strgar and Harwath, 2023)。

如表2所示，本文所提方法在所有评价指标上优于现有方法，实现了最优性能，证明了序列建模在音素分割任务上的有效性。segfeat和superseg两种方法分别依赖于MFCC和对数梅尔频谱图进行特征提取，尽管这两种方法均采取了额外策略以约束预测结果，但传统特征提取方法在捕获音素级别的细节信息及其上下文信息方面存在限制，导致了性能的局限。wav2vec 2.0 finetune方法首次将自监督预训练模型应用于音素分割任务，并取得了一定成就，但其微调策略主要基于对独立音频帧的预测与优化，未能充分考虑音频序列的时序依赖性和标签间的转移关

[1]https://github.com/facebookresearch/fairseq/tree/main/examples/hubert


（c）2024 中国中文信息学会计算语言学专业委员会

系，从而限制了模型的整体性能。本文方法在严格评价指标（*）的性能提升更加明显，充分体现了从序列角度优化并预测的合理性。

| 数据集 | 方法 | 标准指标 | | | | 严格指标 | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 准确率 | 召回率 | F1 | R-val | 准确率* | 召回率* | F1* | R-val* |
| TIMIT | segfeat | 94.03 | 90.46 | 92.22 | 92.79 | - | - | - | - |
| | superseg | 95.63 | 94.77 | 95.29 | 95.82 | - | - | - | - |
| | W2V2 finetune | 96.90 | 96.30 | 96.60 | 97.04 | 94.35 | 93.91 | 94.13 | 94.96 |
| | 本文方法 | **97.47** | **97.51** | **97.49** | **97.86** | **95.77** | **95.18** | **95.48** | **96.08** |
| Buckeye | segfeat | 85.40 | 89.12 | 87.23 | 88.76 | - | - | - | - |
| | superseg | 89.92 | 89.94 | 89.93 | 91.40 | - | - | - | - |
| | W2V2 finetune | 94.01 | 93.08 | 93.54 | 94.41 | 90.56 | 90.28 | 90.42 | 91.81 |
| | 本文方法 | **94.98** | **95.20** | **95.09** | **95.82** | **92.73** | **92.98** | **92.85** | **93.91** |

表2. 在TIMIT和Buckeye数据集上与其它方法的比较，*代表严格指标，-代表未提供分数

#### 4.2.2 声学上下文特征序列对音素分割性能的影响

CRF的解码结果依赖于其输入的发射得分，而不同的声学上下文特征序列所产生的得分也有所不同，为了评估不同特征序列对音素分割任务性能的影响，我们设置了多组实验来探究最适合CRF层的输入嵌入，具体实验设置如下：

（1）MFCC+BiLSTM：使用标准的MFCC特征提取方法，结合BiLSTM捕获音频的上下文依赖性。

（2）Whisper+BiLSTM：使用在大量数据上进行弱监督训练的Whisper(Radford et al., 2023)编码器提取特征，再结合BiLSTM，此处选择预训练的base版本的Whisper编码器[2]。

（3）HuBERT+BiLSTM：采用预训练的HuBERT模型作为特征提取器，利用其最后一层的输出作为音频表征，并结合BiLSTM以捕捉序列的上下文信息。在此配置下，HuBERT在整个训练过程中保持冻结状态。

（4）wav2vec 2.0+BiLSTM finetune：使用预训练的wav2vec 2.0替代HuBERT，并对其进行微调，为了公平，同样选择基于Librispeech 960小时进行训练的预训练检查点[3]。

（5）HuBERT finetune：直接使用HuBERT输出的表征计算发射得分，训练过程中对HuBERT进行微调，无BiLSTM层。

（6）HuBERT+LSTM finetune：将BiLSTM替换为单向LSTM，并对HuBERT进行微调。

（7）HuBERT+Transformer finetune：使用基于自注意力机制的Transformer编码器代替BiLSTM。Transformer在处理序列数据时捕捉长距离依赖，同时通过多头注意力机制并行处理不同表示子空间中的信息，设置隐藏层维度为768，多头注意力头数为8，训练过程中对HuBERT进行微调。

（8）HuBERT+BiLSTM finetune：本文采用的方法，结合HuBERT和BiLSTM，并对HuBERT进行微调。

实验结果如表3所示，比较（1）、（2）与其他结果，发现无论是否进行微调，使用自监督预训练模型（HuBERT或wav2vec 2.0）均优于传统的MFCC特征提取方法，这证实了自监督模型在捕获音素级信息及其上下文关系方面的优势，同时我们发现使用Whisper编码器提取特征在音素分割任务上表现不佳，这或许是因为Whisper主要适用于语音识别、语音翻译等任务，在音素粒度的表征较差。对比（3）、（8）可发现，对HuBERT进行微调能够明显提高其表征能力，比冻结预训练模型的方法能取得更佳的效果。实验（4）和（8）的比较可以证明在同样的条件下HuBERT比wav2vec 2.0在本任务上的效果略好，这和(Strgar and Harwath, 2023)的结论不同，这一发现可能归因于HuBERT和wav2vec 2.0不同的预训练策略所带来的上下文表征能力差异。相比于wav2vec 2.0通过计算量化器得到的码本与Transformer编码器输出的隐层向量之间的对比损失来优化模型，HuBERT通过离线聚类和掩蔽预测过程的交替进行，可以获得更优

[2]https://github.com/openai/whisper
[3]https://github.com/facebookresearch/fairseq/tree/main/examples/wav2vec


(c) 2024 中国中文信息学会计算语言学专业委员会

的上下文表征能力。通过CRF损失以全局序列建模并微调HuBERT更能激发其表征效果，使其在本任务上性能超越了wav2vec 2.0。（5）和（8）的对比实验证明了BiLSTM优化序列上下文依赖关系的有效性，进一步对比（6）和（8）验证了BiLSTM比单向LSTM取得更好的效果，证明了音素分割任务受到前后序列的双重影响。（7）和（8）的对比发现Transformer编码器在本任务上的结果略微低于BiLSTM，考虑到本任务的音频序列较短，同时相邻帧的依赖关系比全局的依赖关系更加重要，这也导致了Transformer编码器无法发挥其捕捉长距离依赖关系的优势。

| 嵌入 | F1 | R-val | F1* | R-val* |
|---|---|---|---|---|
| (1) | 94.23 | 95.03 | 90.22 | 91.64 |
| (2) | 52.50 | 56.71 | 48.50 | 55.22 |
| (3) | 97.11 | 97.47 | 94.33 | 95.04 |
| (4) | 97.39 | 97.78 | 95.19 | 95.87 |
| (5) | 97.36 | 97.75 | 95.18 | 95.86 |
| (6) | 97.41 | 97.78 | 95.35 | 95.80 |
| (7) | 97.32 | 97.72 | 95.10 | 95.78 |
| (8) | **97.49** | **97.86** | **95.48** | **96.08** |

表3. 不同特征输入序列在TIMIT数据集上的实验结果比较，*代表严格指标

#### 4.2.3 损失函数对音素分割性能的影响

为了证明序列建模方法相比于独立预测方法的优势，我们设计了不同的损失函数作为优化目标，在对HuBERT或wav2vec 2.0进行微调并加上BiLSTM框架的基础上，我们探索了以下三种损失函数对实验性能的影响：

（1）交叉熵损失（BCE loss）：通过独立比较每一帧音频与其真实标签间的差异，进行损失计算，其具体计算公式如式（8）。

$$L_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)] \tag{8}$$

（2）焦点损失（focal loss）：对标准的交叉熵损失进行改进，引入了一个系数因子，允许根据预测的准确度自适应调整每个样本对总损失的贡献度，具体计算公式如式（9）。其中$\alpha$是权衡因子，用于减少类别不平衡的影响，而$\gamma$是一个调节参数，用于调整易分类样本的损失贡献。我们设定$\alpha$为0.6，$\gamma$为2，以提高对难以分类的音素边界样本对损失的影响。

$$L_{focal} = -\frac{1}{N}\sum_{i=1}^{N}[\alpha(1-\hat{y}_i)^{\gamma} y_i \log(\hat{y}_i) + (1-\alpha)\hat{y}_i^{\gamma}(1-y_i)\log(1-\hat{y}_i)] \tag{9}$$

（3）CRF损失（CRF loss）：本文采用的方法，通过从整个序列的视角对模型进行建模和优化，其具体计算公式如式（7）。

实验结果如表4所示，焦点损失提升了对难分类样本以及音素边界样本在损失中的惩罚，我们实验中测试了不同的$\alpha$与$\gamma$值，但发现无法产生明显的提升，说明了单独预测的方法在根本上存在局限性。采用CRF损失函数从序列整体进行建模与优化，相较于单个样本的独立预测方法，在所有评价指标上均获得了显著提升。这不仅证明了CRF损失函数在处理音素分割任务时的有效性，也突显了考虑音频序列时序依赖性的重要性。与单独对每一帧进行评估的交叉熵损失和尝试通过调整损失权重来提高难分样本影响的焦点损失相比，CRF损失通过全局优化整个序列的标注，有效捕获了音素边界之间的相互依赖关系，显著优化了模型对于音素分割的理解和预测能力。进一步比较HuBERT与wav2vec 2.0两种预训练模型在不同损失下的结果，我们发现HuBERT在损失函数为交叉熵损失和焦点损失时均落后于wav2vec 2.0，但采用CRF损失后指标的提升更明显，能够达到最佳的性能，说明了HuBERT通过CRF损失进行序列建模能够高度激发其上下文表征能力。这些发现强调了音素分割序列任务的本质，并且采用合适的损失函数对于提升模型性能具有关键作用。


(c) 2024 中国中文信息学会计算语言学专业委员会

| 嵌入 | 损失函数 | F1 | R-val | F1* | R-val* |
|---|---|---|---|---|---|
| W2V2 + BiLSTM | BCE loss | 96.67 | 97.16 | 93.87 | 94.77 |
| | focal loss | 96.74 | 97.43 | 93.89 | 94.92 |
| | CRF loss | 97.39 | 97.78 | 95.19 | 95.87 |
| HuBERT + BiLSTM | BCE loss | 96.53 | 96.96 | 93.70 | 94.57 |
| | focal loss | 96.70 | 97.10 | 93.68 | 94.61 |
| | CRF loss | **97.49** | **97.86** | **95.48** | **96.08** |

表4. 不同损失在TIMIT数据集上的实验结果比较，*代表严格指标

此外，音素分割可能会应用于一些实时场景，为了验证本文方法引入复杂的解码方式后的实时性，我们进一步比较了不同损失函数的解码效率。具体而言，我们在TIMIT测试集上比较了不同方法的RTF值（RTF=解码时间/音频持续时间），各组实验均采用微调HuBERT模型并结合BiLSTM，只有损失函数不同，结果如表5所示，我们的模型在解码时间上确实会多于独立预测的方法，但是RTF值也接近于1，表示能够在一段音频的持续时间结束时完成解码，在实时应用的场景下并不会造成明显的效率降低。

| 损失函数 | RTF |
|---|---|
| BCE loss | 0.761 |
| focal loss | 0.776 |
| CRF loss | 1.093 |

表5. 不同损失在TIMIT数据集上的解码效率比较

#### 4.2.4 跨语言音素分割实验

本文提出的音素分割方法在训练阶段没有使用任何额外的音素标签作为监督信息，同样，在推理阶段也完全摒弃了对音素词典的依赖。这种独立性不仅简化了模型的训练和应用过程，而且理论上模型具备跨语言迁移的强大潜力。

为了深入探索并验证本文方法的跨语言应用能力，我们在老挝语测试集上进行了测试。具体来说我们在英语数据集TIMIT上训练模型，然后将训练好的模型直接应用于老挝语的测试。值得注意的是，老挝语属于声调语言，其声调的多样性高于英语，在语言学特性上和英语有着显著差异，这对跨语言迁移产生了严峻的考验。

在没有对模型进行任何针对老挝语的微调操作的前提下，我们对比了本文方法与现有最佳模型在老挝语上的表现。实验结果如表6所示：由于在训练中未曾接触老挝语语料，模型在老挝语的测试中性能确实有所下降，但本文提出的模型相比于现有模型，仍展现出了更为卓越的性能，达到了71.07的R-value。这一结果表明了本文方法在处理未知语言时，仍然能够保持一定的鲁棒性和适应性。

| 方法 | 准确率 | 召回率 | F1 | R-val |
|---|---|---|---|---|
| W2V2 finetune | 64.65 | 70.99 | 67.67 | 70.97 |
| 本文方法 | **66.89** | **82.33** | **73.81** | **71.07** |

表6. 在老挝语测试集上的实验结果比较

对于跨语种泛化能力的来源，我们认为传统的独立预测的方法只能学习到音频帧与音素边界的对应关系，即模型只能掌握如何根据当前帧的特征判断是否为音素边界，而我们提出的序列建模的方法不仅能够学习到这一对应关系，还能学习到相邻标签之间的依赖以及整个序列的长期依赖。对于不同的语言，音素类别会存在较大差异，传统方法在面对新的语言时可能会遇到很多训练中未见过的特征，从而导致性能急剧下降。我们提出的模型虽然也会面临这一问

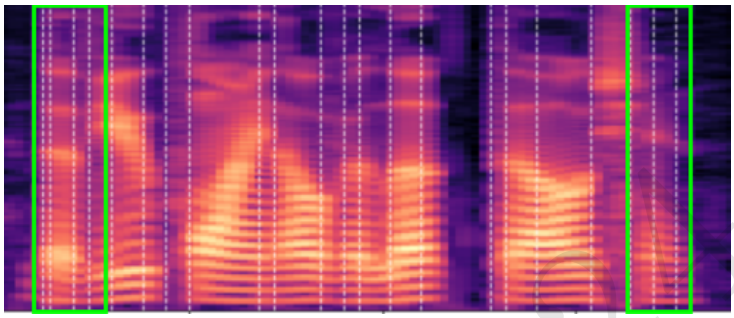

(c) 2024 中国中文信息学会计算语言学专业委员会

题，但我们的模型相比传统方法学习到更多的序列间的依赖关系，并且对于不同语言，相邻标签之间的依赖与整体序列的依赖上的差异会相比音素特征的差异更小，因此本文方法在跨语言实验上展现了更好的能力。

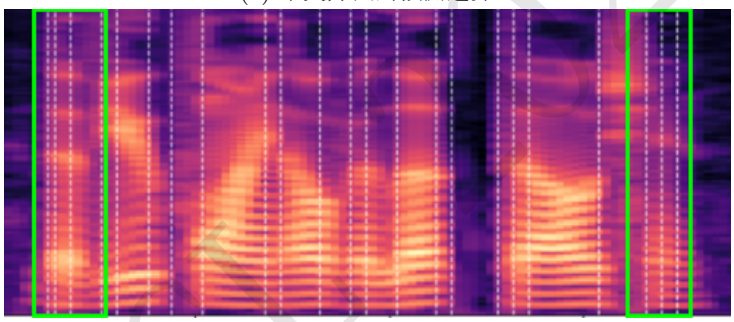
我们采用的预训练HuBERT模型仅在英文数据上进行了预训练，可能无法准确捕捉到老挝语的特定声学特征。然而，与基于独立预测的模型相比，本文方法通过序列建模，能够更有效地利用音频序列中的时序依赖信息。正是由于对音频序列时序特性的深度挖掘和利用，使得本文方法在跨语言迁移的应用上相比现有技术取得了更优的性能表现。

#### 4.2.5 分割实例分析

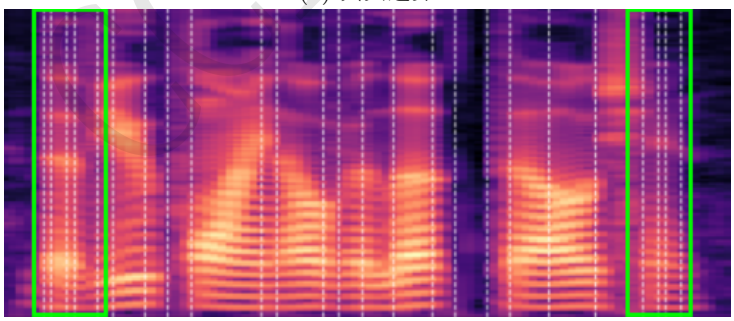
为了更加直观地分析本文方法的效果，我们选取了TIMIT数据集中的单条音频进行分割效果的展示。图3展示了一段特定音频的原始频谱图以及其中的音素边界，边界处为图中的白色虚线。其中（a）为本文方法的预测边界，（b）为真实标签中的边界，(c)为(Strgar and Harwath, 2023)中的最佳模型的预测边界。



(a) 本文方法的预测边界



(b) 真实边界



(c) 对比方法的预测边界

图3. 音素分割实例

通过对比分析这三个边界实例，可以明显观察到，在图3中我们标注出来的部分，（c）相较于真实边界，产生了若干不必要的预测边界。这种现象表明，尽管现有的方法在评价指标上表现出色，但它在预测时显著缺乏对音频序列时序信息的深入理解和准确捕捉。相反，本文提出的方法与真实的边界结果在大多数情况下保持了高度一致，这一比对结果不仅直观地证实了本文方法的有效性，也突显了它在理解音频序列的时序信息方面的明显优势。

此外，通过之前的实验验证，我们已经明确地证明了本文提出的方法在应用严格的评价指标时，能够取得相比其他现有方法更加显著的性能提升。特别是在解决音素边界的冗余预测问

题上，序列建模方法展现了其独特的有效性。这种成效归功于本文方法对音频数据内在时序依赖性的深刻理解和准确建模，确保了在保持预测精度的同时，有效避免了不必要的预测冗余。

## 5 结论

针对现有音素分割方法在捕获音频序列的时序依赖性和标签间关系方面的局限性，本文提出了一种融合语音自监督预训练模型和序列建模的音素分割策略。通过结合预训练的HuBERT模型和BiLSTM层，以及在序列层面优化的CRF损失函数，本文方法不仅优化了音素边界的识别准确性，还增强了模型对音频数据上下文信息的捕捉能力。实验结果表明，相较于传统的独立预测框架，本文方法在TIMIT和Buckeye数据集上均展现出了更优的性能，证明了序列建模对于提高音素分割任务准确性的重要作用。未来的工作将继续探索自监督学习和序列建模在音素分割及其他语音分割任务中的应用。此外，考虑到自监督预训练模型的强大表征能力，未来的研究还将探讨这类模型在跨语言和跨任务场景下的迁移学习能力，以期达到更广泛的应用和更优的性能。

## 参考文献

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518.

Alex Graves and Alex Graves. 2012. Long short-term memory. In *Supervised sequence labelling with recurrent neural networks*, pages 37–45. Springer.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.

Ankita Pasad, Bowen Shi, and Karen Livescu. 2023. Comparative layer-wise analysis of self-supervised speech models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Binghuai Lin and Liyuan Wang. 2022. Learning acoustic frame labeling for phoneme segmentation with regularized attention mechanism. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7882–7886.

Dac-Thang Hoang and Hsiao-Chuan Wang. 2015. Blind phone segmentation based on spectral change detection using Legendre polynomial approximation. *The Journal of the Acoustical Society of America*, 137(2):797–805.

David Rybach, Christian Gollan, Ralf Schluter, and Hermann Ney. 2009. Audio segmentation for speech recognition using segment features. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4197–4200.

Felix Kreuk, Joseph Keshet, and Yossi Adi. 2020. Self-supervised contrastive learning for unsupervised phoneme segmentation. *arXiv preprint arXiv:2007.13465*.

Felix Kreuk, Yaniv Sheena, Joseph Keshet, and Yossi Adi. 2020. Phoneme boundary detection using learnable segmental features. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8089–8093.

Hyeongju Kim and Hyeong-Seok Choi. 2023. Towards trustworthy phoneme boundary detection with autoregressive model and improved evaluation metric. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Jan Chorowski, Grzegorz Ciesielski, Jarosław Dzikowski, Adrian Łańcucki, Ricard Marxer, Mateusz Opala, Piotr Pusz, Paweł Rychlikowski, and Michał Stypułkowski. 2021. Aligned contrastive predictive coding. *arXiv preprint arXiv:2104.11946*.

Jian Zhu, Cong Zhang, and David Jurgens. 2022. Phone-to-audio alignment without text: A semi-supervised approach. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8167–8171.


(c) 2024 中国中文信息学会计算语言学专业委员会

Joerg Franke, Markus Mueller, Fatima Hamlaoui, Sebastian Stueker, and Alex Waibel. 2016. Phoneme boundary detection using deep bidirectional lstms. In *Speech Communication; 12. ITG Symposium*, pages 1–5.

Joey Tianyi Zhou, Hao Zhang, Di Jin, Xi Peng, Yang Xiao, and Zhiguo Cao. 2019. Roseq: Robust sequence labeling. *IEEE Transactions on Neural Networks and Learning Systems*, 31(7):2304–2314.

John Lafferty, Andrew McCallum, Fernando Pereira, and others. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, volume 1, number 2, page 3. Williamstown, MA.

John S Garofolo. 1993. TIMIT acoustic phonetic continuous speech corpus. *Linguistic Data Consortium, 1993*.

Joseph Keshet, David Grangier, and Samy Bengio. 2009. Discriminative keyword spotting. *Speech Communication*, 51(4):317–329.

Joseph Keshet, Shai Shalev-Shwartz, Yoram Singer, and Dan Chazan. 2005. Phoneme alignment based on discriminative learning. In *INTERSPEECH*, pages 2961–2964.

Luke Strgar and David Harwath. 2023. Phoneme segmentation using self-supervised speech models. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 1067–1073.

Mark A Pitt, Keith Johnson, Elizabeth Hume, Scott Kiesling, and William Raymond. 2005. The Buckeye corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication*, 45(1):89–95.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, pages 498–502.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. FAIRSEQ: A Fast, Extensible Toolkit for Sequence Modeling. *NAACL HLT 2019*, page 48.

Okko Johannes Räsanen, Unto Kalervo Laine, and Toomas Altosaar. 2009. An improved speech segmentation quality measure: the r-value. In *Tenth Annual Conference of the International Speech Communication Association*.

Okko Räsanen. 2007. Speech segmentation and clustering methods for a new speech recognition architecture. *Helsinki University of Technology*.

Paul Michel, Okko Räsanen, Roland Thiolliere, and Emmanuel Dupoux. 2016. Blind phoneme segmentation with temporal prediction errors. *arXiv preprint arXiv:1608.00508*.

Sang-Hoon Lee, Ji-Hoon Kim, Hyunseung Chung, and Seong-Whan Lee. 2021. Voicemixer: Adversarial voice style mixup. *Advances in Neural Information Processing Systems*, 34:294–308.

Santiago Cuervo, Adrian Lancucki, Ricard Marxer, Paweł Rychlikowski, and Jan K Chorowski. 2022. Variable-rate hierarchical cpc leads to acoustic unit discovery in speech. *Advances in Neural Information Processing Systems*, 35:34995–35006.

Tusarkanta Dalai, Tapas Kumar Mishra, and Pankaj K Sa. 2023. Part-of-speech tagging of Odia language using statistical and deep learning based approaches. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(6):1–24.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

CCL 2024

# 近三十年域外汉籍研究的现状与展望——基于文献计量分析和知识图谱绘制

**唐榕骏**[1], **彭志峰**[*1,2,3]
[1] 暨南大学文学院 [2] 暨南大学汉语方言研究中心
[3] 广东省岭南数字人文实验教学示范中心
{moniyi}@yeah.net, {pengzhifeng}@jnu.edu.cn

## 摘要

域外汉籍研究对中华文化传承传播意义重大、成果丰富，但缺少文献计量分析。本研究的目的是通过CNKI优质期刊论文，分析中国近30年域外汉籍研究的演进趋势。通过知识图谱、普赖斯理论、Leydesdorff引文理论、斯皮尔曼等级相关系数与关键词共现、聚类、突现、时线分析研究学科演进趋势。研究发现，“汉籍整理与数字化研究”“文化交流与文化传播”为前沿热点，揭示出领域内研究工作已从回归普查转向应用转化。本研究为学科发展提供了参考数据。

**关键词：** 域外汉籍；知识图谱；文献计量；共现分析

# Extraterritorial Chinese Texts in the Last Thirty Years: Research Advances and Future Perspectives Based on Knowledge Mapping and Bibliometric Analysis

**Tang rongjun**[1], **Peng zhifeng**[1,2,3]
[1] College of Arts, Jinan University [2] Institute of Chinese Dialects, Jinan University
[3] Provincial Demonstration Center for Lingnan Experimental Digital Humanities Education
{moniyi}@yeah.net, {pengzhifeng}@jnu.edu.cn

## Abstract

The study of Extraterritorial Chinese Books (ECBs) holds great significance for the preservation and dissemination of Chinese culture, however, there is a dearth of bibliometric analyses in this field. This study aimed to investigate the evolutionary trends of ECBs research in China over the past 30 years. Base on CNKI database, the research advances of ECBs were analyzed using knowledge graph visualization, Price's law, Leydesdorff's citation theory,and Spearman's rank correlation coefficient, along with keyword co-occurrence analysis, clustering, and timeline analysis. The hotspots identified include “ECBs arrangement and digitization” and “cultural exchange and dissemination”. It suggests that ECBs research has evolved from a focus on cataloging and inventory-taking towards the application. This study informs the future development of the discipline.

**Keywords:** Extraterritorial Chinese books , knowledge map , bibliometrics , co-occurrence analysis

*通讯作者：彭志峰（pengzhifeng@jnu.edu.cn）

第二十三届中国计算语言学大会论文集，第637页-第649页，太原，中国，2024年7月25日至28日。
卷1：主会论文
(c) 2024 中国中文信息学会计算语言学专业委员会

# 1 引言

域外汉籍指域外所藏中国古籍、域外刻本及域外人士撰写的汉文典籍(金程宇, 2010)。历史上，我国大量珍稀古籍流入海外。新中国以来，我国政府部门与学术机构高度重视域外汉籍的回归和研究工作。值得关注的是，相关政策重点发生显著的历时迁移：“十一五”时期《国务院办公厅关于进一步加强古籍保护工作的意见》提出“大量珍贵古籍流失海外，加强古籍保护刻不容缓”，“十三五”时期《“十三五”时期全国古籍保护工作规划》以“建立海外中华古籍回归工作机制，有计划、分步骤开展海外中华古籍回归工作”为目标，2020年《国家新闻出版署关于编制“十四五”国家重点图书、音像、电子出版物出版规划和2021—2030年国家古籍规划的通知》倡导“对存藏海外中国古籍进行系统性整理，加强古籍数字化开发利用”，2022年《关于推进新时代古籍工作的意见》鼓励“充分利用海外文化平台开展古籍对外宣传推广活动”。

中国域外汉籍研究始于1986年“中国域外汉籍国际学术会议”(张伯伟, 2006)，已有三十余年历史，成果丰富。但目前该领域的研究综述主要依托短时段样本开展定性研究(奎欣怡, 2023)，缺乏基于历时数据的量化客观分析，如傅荣贤、华建铭和李琳分别综述了1949年至2019年中国古典文献学(傅荣贤and 华建铭, 2020)和古代文学(李琳, 2014)的发展的演变历程，其中片段式地简要概述了域外汉籍研究。金程宇、刘泰廷分别综述了2000-2009年(金程宇, 2010)、2010-2014年(刘泰廷, 2017)阶段的域外汉籍学术研究成果，却未对域外汉籍研究内部的发展沿革进行量化梳理和剖析。因此，该领域目前面临着政策转向显示出的社会需求升级，与学术研究全景的未经量化、不可捉摸现状之间的信息不对等。这不利于学者把握研究方向和规划发展路径。因此，本研究目的是系统分析中国知网（CNKI）近三十年域外汉籍研究优质期刊（含辑刊）论文数据，总结归纳出发展沿革与演进趋势。为此，拟通过定量与定性结合的方式，基于知识图谱绘制和文献计量综合分析筛选的809篇文献。这将为本领域的未来研究方向提供依据。

# 2 研究方法

## 2.1 发文信息分析理论与算法

### 2.1.1 普赖斯曲线

普赖斯曲线认为科学文献的增长与时间成指数函数关系。其公式为：$F(t) = a \times ebt$，F(t)表示预测发文量，a为初始年发表量，e为常数，b为时间常数，表示期刊持续增长率。b值的计算公式为$\mathrm{b} = \sqrt[n]{\frac{A}{a}} - 1$，其中n为统计年限，A为发文总量。本研究用于拟合实际发文情况，分析发文时序发展健康程度。

### 2.1.2 平均值、方差、标准差

发文量机构年发文量的平均值、方差、标准差，可以用于评估机构发文量的变异程度。其公式依次为：

$$\text{Mean} == \frac{\sum \text{number of publications}}{\text{number of institution}}$$

$$\text{Variance} = \frac{\sum(\text{number of publications} - \text{Mean})^2}{\text{number of institution}}$$

$$\text{Standard Deviation} = \text{sqrt(Variance)}$$

### 2.1.3 斯皮尔曼等级相关系数

斯皮尔曼等级相关系数是用于衡量两个变量相关性的非参数统计方法(白雪涵, 2022)，本研究使用其验证刊文量前20期刊与CNKI影响因子高低的相关度。其中，$\rho$表示斯皮尔曼相关系数，$d_i$表示两个变量在排序中的差异，n表示样本数量。具体公式为：$\rho = 1 - 6\sum d_i^2/n\left(n^2-1\right)$。

### 2.1.4 普赖斯核心作者定律与综合指数公式

普赖斯定律公式用于计算核心作者范围。公式为$\mathrm{m} = 0.749 * (\mathrm{n_{max}})^{1/2}$,其中$\mathrm{n_{max}}$为发文最多作者的总发文量(丁学东, 1993)。确认范围后，进一步通过综合指数法测定核心作者，其中Zi是第i位候选人综合指数值，$\mathrm{X_i}$，$\mathrm{Y_i}$分别是第i位候选人的发文量、被引量，X，Y分别是候选人平均发文量、平均被引量。公式为$\mathrm{Zi} = \mathrm{Xi/X} * 100 * 0.5 + \mathrm{Yi/Y} * 100 * 0.5$。

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

### 2.2 研究热点前沿分析理论与算法

#### 2.2.1 Leydesdorff长期、短期引用理论

引文分析依据Leydesdorff提出的长期引用、短期引用理论(Loet et al., 2016)，其认为文献根据发表到成为高被引文献的时间长短，可以分为长期、短期引用文献。短期引用说明被引文献处于研究前沿；长期引用说明被引文献对于学科基础理论的形成作出贡献。

#### 2.2.2 高频词汇边界算法

高频关键词根据高频词边界公式为$T = (-1 + \sqrt{1 + 8I}) \div 2$确定(Donohue, 2014)。T为高频关键词个数，I为频次为1的关键词个数。

#### 2.2.3 共现分析

共现分析的基本原理是：两两统计一组词在文献中的出现频次，通过共现次数测度其亲疏关系。共现次数多说明这组词关系更密切，语义上更易产生联系(李杰and 陈超美, 2016)。分析过程如图1所示。

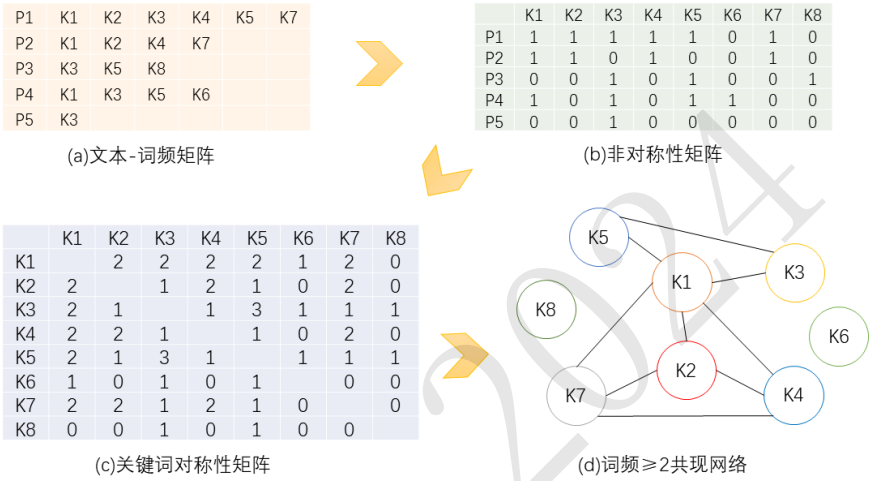



图 1: 共现分析原理示意图

### 2.3 数据来源

“域外汉籍”简称“汉籍”，为避免以“域外汉籍”“海外汉籍”为主题检索时有遗漏，或以“汉籍”为主题检索时结果宽泛，采取“系统+人工”方式筛选样本。首先，在CNKI以“主题=汉籍”进行检索，“期刊来源类别=北大核心+CSSCI+CSCD+AMI+SCI+EI”以保障论文优质，并纳入两本域外汉籍专门辑刊中的相关研究文献。时间范围为1992年至2023年。检索时间为2024年3月1日。其次，人工去除“西学汉籍”“汉族典籍”“汉代典籍”“英译汉典”等无关文献及干扰信息。共得到809篇样本文献。

### 2.4 研究流程

本研究基于CNKI采集文献数据、CiteSpace与VOSviewer绘制知识图谱、Microsoft Excel绘制图表、Note Express辅助词频分析。具体流程如图2所示。

## 3 发文信息分析

### 3.1 发文时间分布与拟合分析

论文数时序变化是衡量领域发展的重要指标(邱均平et al., 2012)。如图3所示，年发文量时间分布整体呈上升趋势。按照年发文量与发文增速阶段差异，可将其分为“萌芽期”（1992-2005）、“成长期”（2006-2012）和“高速期”（2013-2023）。其年均发文量依次为10篇以下，20篇以下，20篇以上。为研究发文时间分布的健康程度，分别绘制发文量曲线和普赖斯曲线。结果显示，“萌芽期”两条曲线拟合度高，但此后年发文量低于普赖斯曲线，


(c) 2024 中国中文信息学会计算语言学专业委员会

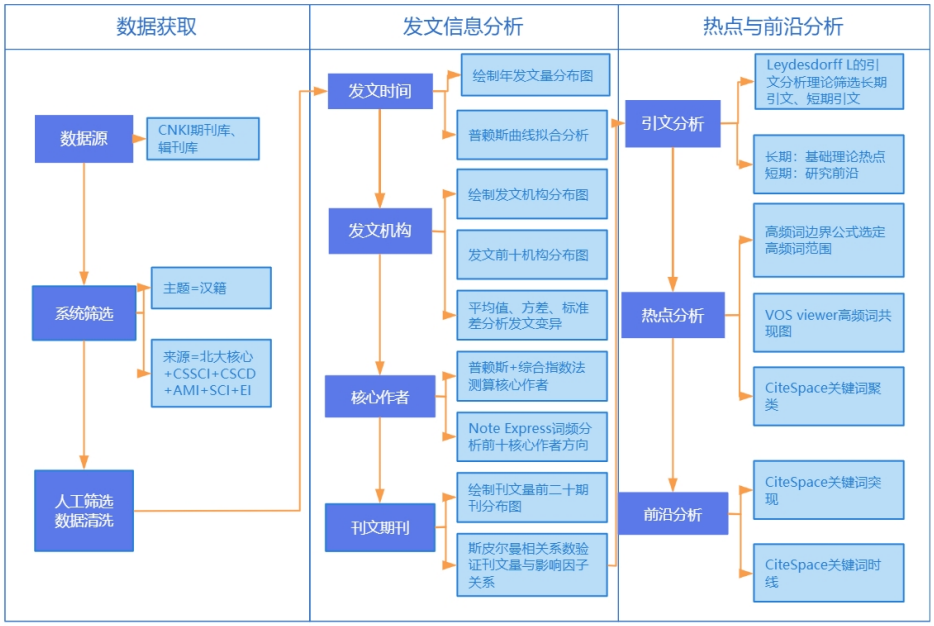



图 2: 研究路线流程图

这说明该领域学科发展的干扰因素可能较多。分析发文峰谷年份的发文环境可能有助于了解干扰因素。对关键峰值2006年分析发现，2005年创办了《域外汉籍研究集刊》，2006年《学习与探索》开设域外汉籍专栏，说明专业刊物对学科的积极影响。此外，2020年出现明显发文量低谷，这可能是受新冠疫情影响引发的线下交流与调查困难导致。但情况在一年后快速好转，说明该领域已形成一定的抗风险能力。

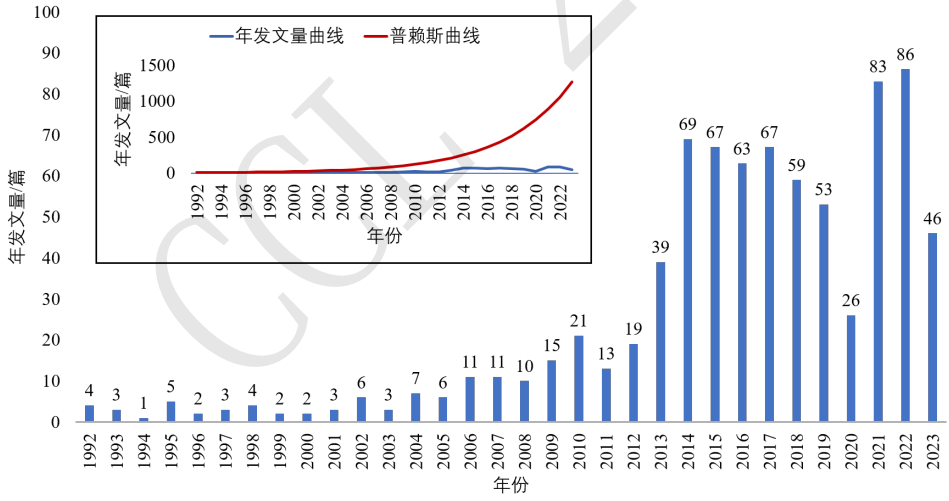



图 3: 域外汉籍研究发文时间分布与拟合曲线

### 3.2 研究机构分布

分析发现，域外汉籍研究机构以文学和历史院系为主（图4a）。提取机构信息，按照机构一级单位归并编码，并生成发文量前十机构逐年发文量图（图5）。统计发现，发文最多的机构是南京大学（89篇），其次是山东大学（28篇）、复旦大学（28篇）。高产出机构大部分为“双一流”高校。南京大学于2000年2月成立了域外汉籍研究所，山东大学2013年起承担文化部“全球汉籍合璧工程”，复旦大学2022年获批国家社科基金重大项目“东亚汉籍版本学史”。上述因素可能对其成果产出有增益。在“双非”高校中，云南省蒙自市红河学院发文量为全国第11（9篇），


(c) 2024 中国中文信息学会计算语言学专业委员会

其在2008年成立越南研究中心。全球汉籍分布GIS系统[0]统计，与我国临近的越南等国汉籍资源丰富（图4b）。红河学院的高产出，说明边境省份高校开展域外汉籍研究地缘优势显著。

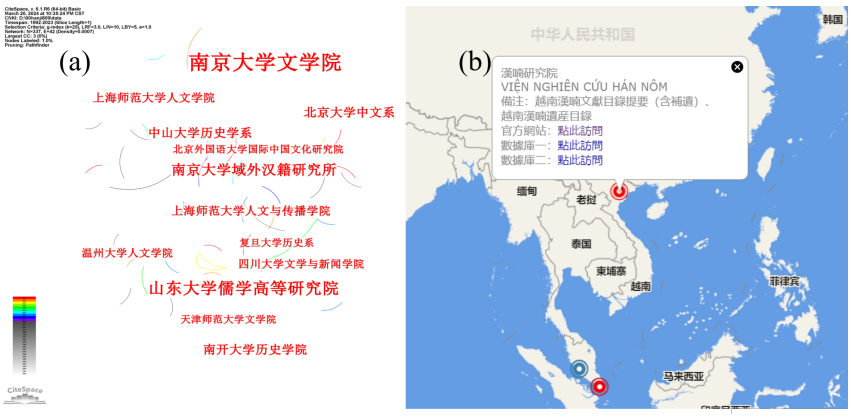



图 4: 发文机构分布与全球汉籍分布GIS系统越南汉籍地图示意

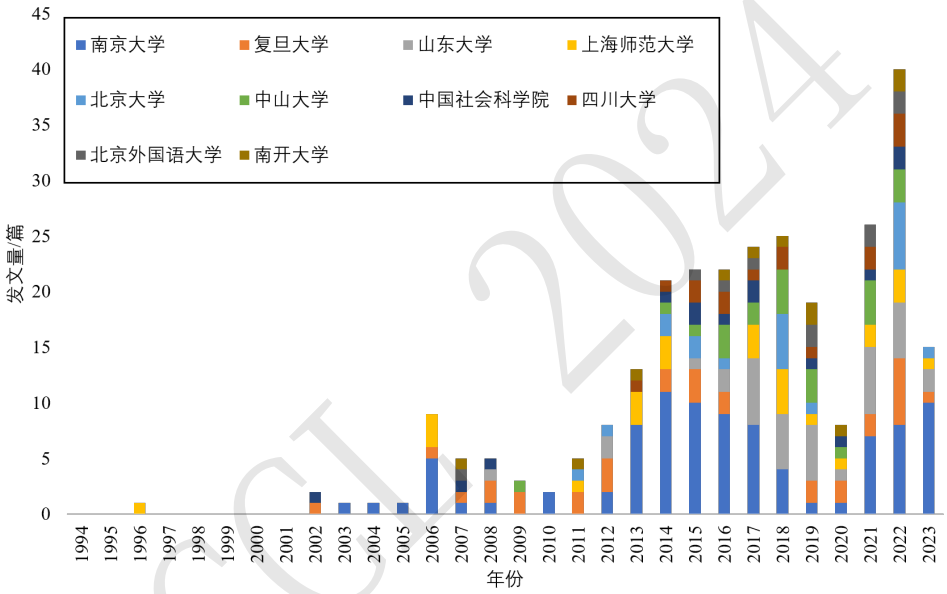



图 5: 发文前十机构逐年发文量

为评估机构发文量的变异程度，进一步计算发文前十机构的年发文量的平均值、方差、标准差（表1），结果发现，2014年标准差最高，为3.11，可知这一年各机构发文量差异最大。且此后标准差整体处于较高水平，说明“高速期”各机构成果产出差异变大。

### 3.3 核心作者分析

分析核心作者及其研究方向，有助于快速把握该领域的研究热点和最新动态。为确定核心作者，通过综合指数法进行测定。首先，使用普赖斯核心作者公式计算m=2.8≈3，说明发文量在3篇及以上的为核心作者候选人。结果显示，366位作者（图6，N=366）中有34位候选人，合计发文159篇，占总成果量的19.6%。其次，将发文量和被引量权重值分别定为0.5和0.5(钟文娟, 2012)，计算综合指数值（附表B）。最后，使用Note Express对综合指数前10的核心作者成果进行词频分析（表2）。结果发现：核心作者关注日本、越南、朝鲜半岛地区，目录书、使臣著作、东亚汉诗类作品，数字化、传播策略相关策略经验。

[0]全球汉籍分布GIS系统.https://www.wenxianxue.cn/gis/.


(c) 2024 中国中文信息学会计算语言学专业委员会

| 年份 | 1994 | 1995 | 1996 | 1997-2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 平均值 | 0 | 0 | 0.1 | 0 | 0.2 | 0.1 | 0.1 | 0.1 | 0.9 | 0.5 | 0.5 | 0.3 | 0.2 |
| 方差 | 0 | 0 | 0.09 | 0 | 0.16 | 0.09 | 0.09 | 0.09 | 2.69 | 0.25 | 0.45 | 0.41 | 0.36 |
| 标准差 | 0 | 0 | 0.3 | 0 | 0.4 | 0.3 | 0.3 | 0.3 | 1.64 | 0.50 | 0.67 | 0.64 | 0.60 |
| 年份 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 | 2021 | 2022 | 2023 |
| 平均值 | 0.5 | 0.8 | 1.3 | 2.1 | 2.2 | 2.2 | 2.4 | 2.5 | 1.9 | 0.8 | 2.6 | 4 | 1.5 |
| 方差 | 0.45 | 1.16 | 5.81 | 9.69 | 7.56 | 5.76 | 6.24 | 4.05 | 1.49 | 0.36 | 5.04 | 4 | 8.45 |
| 标准差 | 0.67 | 1.08 | 2.41 | 3.11 | 2.75 | 2.40 | 2.50 | 2.01 | 1.22 | 0.60 | 2.24 | 2.00 | 2.91 |

表 1: 各机构年发文量平均值、方差、标准差

另外，图6显示，海外高校作者表现积极，如陈翀（日本广岛大学）、水上雅晴（日本中央大学）。核心作者间合作较少，但存在多组跨国合作，如金程宇（南京大学）与衣川贤次（日本花园大学）、河野贵美子（日本早稻田大学）与葛继勇（郑州大学）。




图 6: 作者分布与合作共现图

### 3.4 刊文期刊分布

样本文献分布于161本期刊中，学科分布以文献学、文学、图书情报学为主，说明该领域目前文献辑佚、整理相关成果较为丰富，为后续研究打下坚实基础。期刊主要集中在社会科学领域，说明自然科学学科汉籍研究成果较少。《关于推进新时代古籍工作的意见》专门提出“梳理挖掘古典医籍精华”的要求，但当前研究缺口较大。统计刊文量前二十的期刊（图7），发现刊文最多的期刊为《域外汉籍研究集刊》（394篇），其次为《汉籍与汉学》（43篇），《文献》（20篇），其中前两者为学科专门性辑刊。

斯皮尔曼相关系数可以验证刊文量前二十的期刊与CNKI影响因子高低的相关度大小。将刊文量与影响因子整理为表3（辑刊无影响因子数据，不计入统计），计算发现两组数据的斯皮尔曼相关系数 $\rho$ =-0.4419，表示该领域刊文量前二十期刊与CNKI综合影响因子之间存在一个较弱的负相关关系，说明刊发该领域论文数量的多少与期刊的综合影响因子并没有强烈的直接联系。

### 3.5 域外汉籍研究的热点与前沿

#### 3.5.1 基于引文分析的基础理论与前沿成果

Leydesdorff的引文分析理论将文献根据发文到成为高被引文献时间的长短，分为长期和短期引用。长期引用体现领域基础理论的形成，短期引用体现研究前沿性。将短期引用年限设为5年，分析被引频次前10的高被引文献(胡一尘, 2020)。结果发现，9篇为长期引用，分析其选题方向发现，域外汉籍在东亚汉字文化圈研究范畴已形成基础理论；1篇为短期引用（《中国


(c) 2024 中国中文信息学会计算语言学专业委员会

| 序号 | 姓名 | 高频词 |
|---|---|---|
| 1 | 张伯伟 | 方法；文学；书目；东亚汉文化圈；朝鲜 |
| 2 | 巴兆祥 | 日本；地方志 |
| 3 | 卞东波 | 诗；东亚；日本；诗人 |
| 4 | 金程宇 | 日藏善本书志；中日韩戏剧；东亚汉文小说；东亚汉诗 |
| 5 | 陈翀 | 日本；空海；《日本国见在书目录》；唐诗 |
| 6 | 刘玉 | 越南；使臣；燕行诗 |
| 7 | 王志强 | 日本；朝鲜；越南；《往津日记》 |
| 8 | 毛建军 | 古籍数字化；数据库 |
| 9 | 任晓霏 | 蒙学；教育；传播 |
| 10 | 叶少飞 | 大越史；汉文小说；安南使臣 |

表 2: 前十位核心作者研究方向高频词表

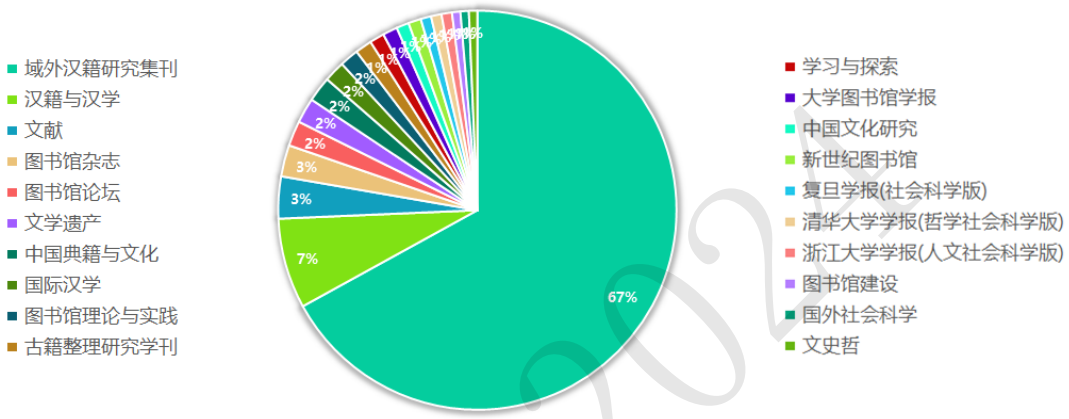



图 7: 刊文量前二十期刊分布图

古代蒙学典籍海外传播和影响研究》），说明“蒙学教育”与“传播策略”是学者近年关注的热点（高被引文章列表可见于附表A）。

### 3.5.2 基于词频、聚类、共现的研究热点分析

关键词的频率、聚类、共现可以分析研究热点。样本文献包含487个关键词（图9，N=487），根据高频词边界公式T=27，因第27个关键词出现频次为3，故将出现频次大于等于3的关键词认定为高频关键词，合计41个。使用VOSviewer生成高频词共现图，与使用CiteSpace生成的聚类图比对分析可以归纳出三个重要研究热点（图8、9）。在描述热点及内部成员发生的时间节点时，结合时线图、突现图进行分析（图10、11）。

**热点一：“图书馆+目录”模式下的汉籍普查研究**

综合观察图8图书馆、古籍目录等共现词，图9聚类#2图书馆与聚类#6目录学，结合突现与时线图分析。结果发现，从1992年起，以图书馆为依托、以馆藏目录为抓手的汉籍普查已是域外汉籍研究的重要热点。研究对象主要分布在欧美和东亚。一般采用实地调查与数据检索结合的研究方式。欧美汉籍普查研究聚焦法国、英国、俄罗斯、梵蒂冈和美国。欧洲汉籍收藏历史悠久且存量丰富。法国国家图书馆、巴黎汉学研究所图书馆与《皇家图书馆写本目录》为关键普查线索(刘蕊and 岑咏芳, 2019)。英国目前所见的14种英国图书馆所编汉籍目录，涉及大英博物馆、牛津大学博德利图书馆等8间图书馆(谢辉, 2015)。俄罗斯汉籍主要藏于俄罗斯国立图书馆等4间图书馆(郑诚, 2023)，但其馆藏汉籍目录尚未出版。梵蒂冈图书馆的重要目录是伯希和所编《梵蒂冈图书馆所藏汉文写本和印本书籍简明目录》及高田时雄《梵蒂冈图书馆所藏汉籍目录补编》(谢辉, 2022)，后者编目采用汉语拼音并标注汉字，更易读。德国、奥地利、瑞典等国图书馆亦存少量汉籍(谢辉, 2016)。北美汉籍收藏始于1867-1869年中美之间的图书交换(杨

| 期刊名 | 文献 | 图书馆杂志 | 图书馆论坛 | 文学遗产 | 中国典籍与文化 | 国际汉学 |
|---|---|---|---|---|---|---|
| 刊文量 | 20 | 15 | 12 | 12 | 12 | 10 |
| 影响因子 | 0.345 | 1.747 | 2.448 | 0.424 | 0.145 | 0.305 |
| 期刊名 | 图书馆理论与实践 | 古籍整理研究学刊 | 学习与探索 | 大学图书馆学报 | 中国文化研究 | 新世纪图书馆 |
| 刊文量 | 9 | 8 | 7 | 7 | 6 | 6 |
| 影响因子 | 1.417 | 0.09 | 1.589 | 2.457 | 0.465 | 1.071 |
| 期刊名 | 复旦学报(社会科学版) | 清华大学学报(哲学社会科学版) | 浙江大学学报(人文社会科学版) | 图书馆建设 | 国外社会科学 | 文史哲 |
| 刊文量 | 5 | 5 | 5 | 4 | 4 | 4 |
| 影响因子 | 1.202 | 1.173 | 1.173 | 2.091 | 1.833 | 0.944 |

表 3: 刊文前二十期刊刊文量与影响因子对照表

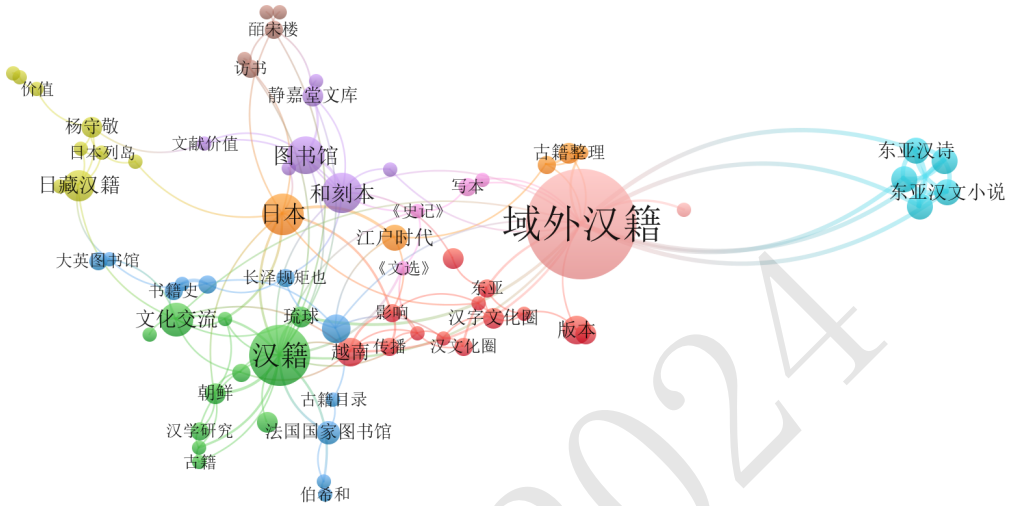



图 8: 高频关键词共现图

海峥, 2022)。主要藏于美国国会图书馆、哈佛大学燕京图书馆等数十间图书馆(严润开, 2001)，其中26间的汉籍普查已完成[1]。东亚汉籍普查研究聚焦于日本、韩国和越南。日本静嘉堂文库和韩国奎章阁热度最高。《"日本所藏中文古籍数据库"目录》《奎章阁图书中国本综合目录》研究最多。越南汉喃研究院和越南国家图书馆从2013年起成为热点（图11）。国内图书馆亦有回流汉籍，如北京北堂图书馆（今西什库教堂附属图书馆）、北京大学图书馆、国家图书馆、天津图书馆和大连图书馆"大谷文库"。虽然海外汉籍普查至今尚无系统明确的统计数据公布，但全球汉籍合璧工程项目组在2023年表示基本已掌握境外汉籍存藏情况，并通过课题方式逐步落实普查工作(陈灏et al., 2023)。因此，"图书馆+目录"模式下的汉籍普查研究，可能在未来几年成果加速产出后结束热点。

**热点二：版本考证与版本流布研究**

综合观察图8和刻本、写本等共现词，图9聚类#1汉籍（聚类成员版本研究、日本本、写本等）与聚类#16汉学（聚类成员安南本、出版等），结合突现与时线图分析。结果发现，版本是域外汉籍研究的热点，关注版本考证和版本流布研究，以1992年和刻本研究为起点，在2016年至2020年热度最高。版本考证研究重视域外中国写本、抄本及孤本。研究对象存藏较为集中的有法国、英国劫购的敦煌遗书。其余孤本存藏分散，仅日藏本研究成果相对较多。如日本所藏稀见明词研究(周明初and 彭志, 2014)。版本流布研究关注域外刻本。其中日本和韩国的"和刻本""朝鲜本"研究热度最高，尤其是江户时代和朝鲜时代。主要是因其汉籍流入时间早，版本丰富，有助于从版本学视角分析"和而不同"的东亚汉文化圈。成果以个案研究为主，尤其关注佛典儒典（#1汉籍）。近十年，越南"安南本"研究热度提升。研究材料获取难、部分文献社会价值不足是该方向的研究难点，可能是其2020年之后前沿性不足的原因（图10）。完整系统的域外汉籍目录与数据库有助于解决该问题。因此，在汉籍普查完成后，该研究可能再次成为前沿热点。

[1]全球汉籍合璧工程.境外藏书机构联系进展情况一览表. https://hanjihebi.sdu.edu.cn/info/1010/1032.htm.


(c) 2024 中国中文信息学会计算语言学专业委员会

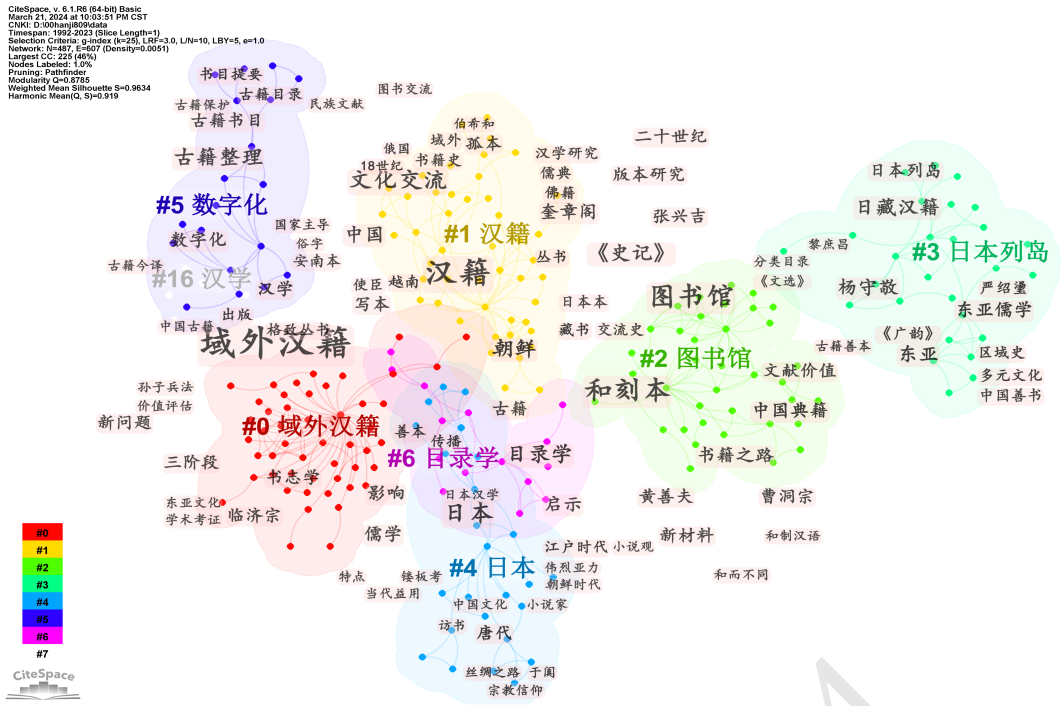



图 9: 关键词聚类图

**热点三：名家名作专题研究**

综合观察图8《史记》《文选》、长泽规矩也等共现词，图9聚类#1汉籍（聚类成员《史记》等）与聚类#4日本列岛（聚类成员严绍璗、《广韵》《文选》等），结合突现与时线图分析。结果发现，域外汉籍名家专题研究中，杨守敬、严绍璗（国内），以及伯希和、岛田翰、长泽规矩也（国外）是研究热点。其中杨守敬研究起步最早，持续最久，为1993年至2019年。严绍璗研究最晚，始于2009年。陆心源、王重民、董康等人的研究热度稍次之，但前沿性更显著，其中对董康的研究热点在2018年至2022年。国内名家研究关注其藏书、访书情况与代表作品。如杨守敬日本访书回传情况、陆心源藏书流失日本的情况、董康日本访书及其《书舶庸谈》、王重民欧洲访书及其所获敦煌遗书和太平天国文献、严绍璗及其《日藏汉籍善本书录》。国外名家研究关注其获取、整理汉籍。如日本学者岛田翰与宋楼藏书去日情况、长泽规矩也中国访书经历及成果、法国学者伯希和在华所购汉籍与海外收藏。名作研究重点关注作品在海外的经典化路径，热点有《广韵》《史记》《文选》《世说新语》，其中《广韵》研究热度开始最早。“世说学”仍在前沿（图11#0），相关研究有日韩对《世说新语》经典化研究、《世说新语》衍生作品研究(王子成, 2023)。总体来看，名家专题研究热点的热度持续时间大都较短，前沿性不显著。名作专题研究热点的热度持续时间较长，且“世说学”研究仍具有一定的前沿性。

### 3.5.3 基于突现、时线的研究前沿与展望

**前沿一：汉籍整理与数字化研究**

基于数字化的汉籍整理出版研究是前沿热点之一，尤其聚焦于数据库平台和数字化技术两方面（图11#5数字化）。数据库平台研究关注数据库建设和应用研究。数字化技术研究关注汉籍数字化整理与转化。“数字化”突现首年是2018年（图10），当前处于研究前沿。

数据库建设研究重点关注日本汉籍数据库。其原因是日本相关建设比较成熟。“日本所藏中文古籍数据库”[2]检索系统是日本目前参与机构最多的汉籍数据库系统。在建设规范和实践路径上有重要参考价值，包括色彩平衡与数字标尺、采集粒度、存储参数与格式、特殊情况处理方面（古籍碎片、无字页、书内夹抄、附属物件）(朱本军, 2020)。欧美汉籍数据库研究热度较低，因其汉籍数据库建设投入不足、发展较慢(张敏and 董强, 2011) 。数据库应用研究起步不

[2]日本所藏中文古籍数据库. http://kanji.zinbun.kyoto-u.ac.jp/kanseki.

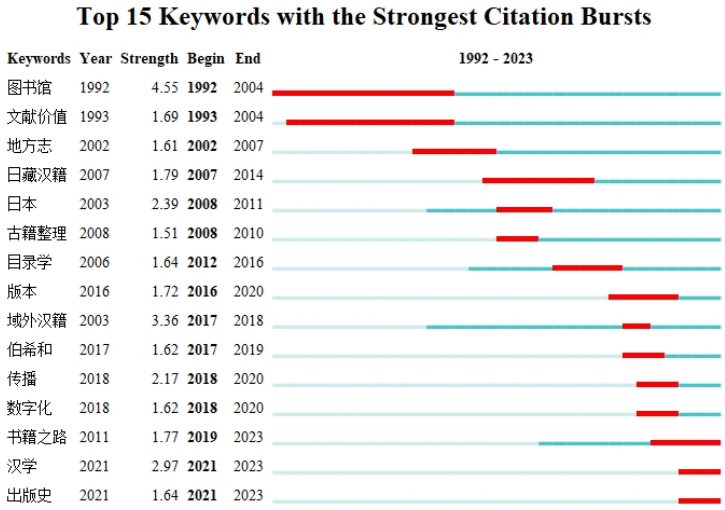



图 10: 域外汉籍研究关键词突现图

久。如2023年《中日汉籍出版中的“鳌头本”》使用国内外数据库检索查考出40余部“鳌头本”(潘超, 2023)。“中国古典文献资源导航系统”[3] 数据库索引为汉籍研究提供便利。但整体来看，汉籍数据库建设仍有不足，大多仍需配合纸本使用。

汉籍数字化研究主要服务于回归汉籍的整理工作。目前汉籍以影像回归方式为主，因此影像采集阶段的清晰度、速度、还原度，以及采集后的影像自适应智能处理技术是影响数字化质量的关键。例如，古籍印章识别技术就是古籍副文本智能处理研究的经典案例(顿泽栋, 2022)。但相关技术研究成果多为影像采集阶段，智能处理阶段成果依然不足(罗渝and 黄璜, 2018)。“全球汉籍合璧”工程等项目显示出，国家主导下汉籍数字回归加速（图11#5），其伴生的汉籍数字化整理工作方兴未艾。更重要的是，数字化整理、加工是汉籍规模化“用起来”的重要前提，因此，本方向的研究在未来较长时间内都将是前沿趋势。

**前沿二：文化交流与传播研究**

汉籍承载着中华文化之精神，“书籍之路”即是中华文化流播之路。“传播”“书籍之路”首现时间为2018年和2019年（图10），图11中多个聚类时线最新的成员词都与传播相关，如#0文化传播、#2唐诗传播、#4传播途径等，说明文化交流与传播研究的前沿性趋势。与汉籍普查、回归、整理等“引回来”研究不同，该方向关注“走出去”，尤其是“再次走出去”的相关研究。包括三类内容：汉籍与文化传播史、汉籍与文化交流融合以及当代传播策略。

汉籍与文化传播史研究，关注汉文化圈国家的汉籍受容。国别上关注日本、韩国、越南；时代上关注中国的唐代、日本的江户时代、朝韩的朝鲜时代。传播途径上关注赏赐（求书）、贸易、劫购、访书等。传播者关注来华使臣、留学僧（生）、贸易商、汉学家。具体研究内容上有蒙学、道教、汉字、儒学、佛学等某一门类汉籍与文化的传播史，以及《三字经》《文心雕龙》《欧苏手简》等某一汉籍及其文化义涵的受容与变容。少量欧美国家汉籍与汉文化传播史研究关注欧洲的18世纪“中国热”时期。汉籍与文化交流融合研究，关注汉籍中外互动研究。研究内容包括“汉籍拟效研究”“融合互鉴研究”。“汉籍拟效研究”有三个方向。其一是海外对汉籍编辑方法拟效，如英国汉学家伟烈亚力以中国传统四分法编著《中国文献解题》。其二是仿写续写研究，如日本《两假名杂字尽》对中国杂字类蒙书的仿作、17至19世纪日本朝鲜笺释摹仿《世说新语》热潮研究。其三是科技典籍仿效应用研究，如中国《农政全书》及其日仿本《农业全书》对日本农业文明与技术发展影响的研究。“融合互鉴研究”关注书籍环流中的多元文化共生，如《纪效新书》等中国武术古籍在朝鲜和日本传播后，其原著及衍生作品回流中国，其作品中展示出的东亚三国文化和而不同的多元融合(蔡艺et al., 2022)。当代传播策略研究，关注如何利用汉籍开展对外交流和文明互鉴。在汉籍译介和学术交流等传统路径外，依托新媒体的汉籍转化传播开始受到关注(姚委委et al., 2021)。但仍属于新兴领域。

[3]中国古典文献资源导航系统.https://www.wenxianxue.cn/view/1.html.


(c) 2024 中国中文信息学会计算语言学专业委员会

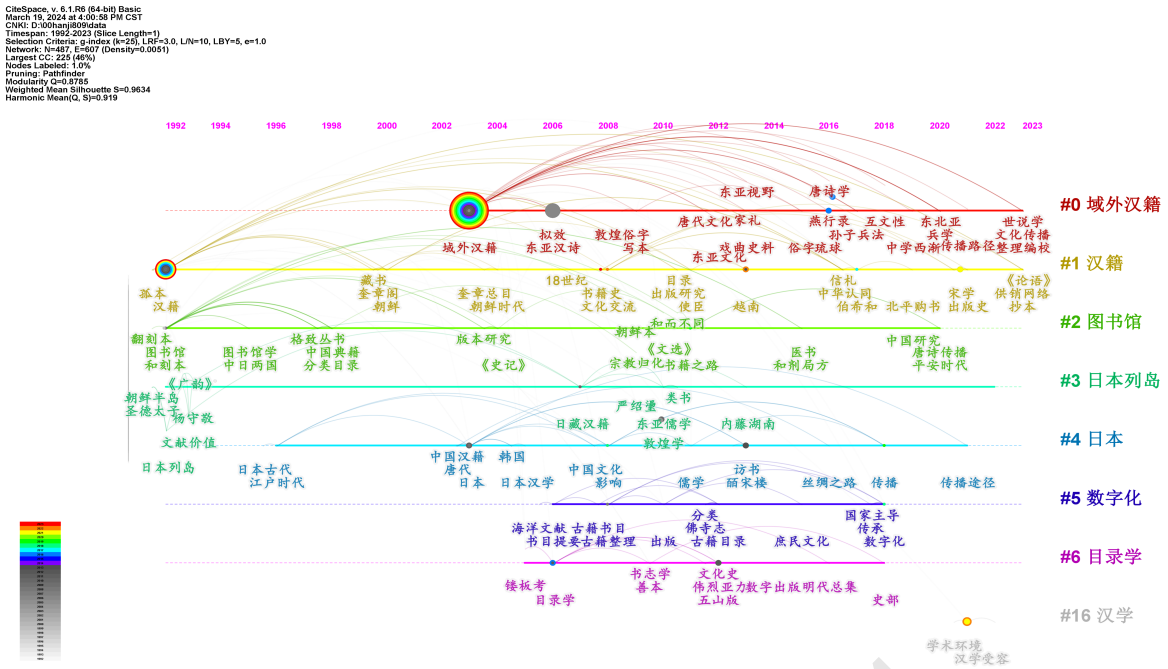



图 11: 域外汉籍研究关键词与时线图

## 4 总结与展望

综上所述，通过定量与定性结合的分析方法，分析以往传统定性研究无法实现的长时段、大数据样本，对域外汉籍研究情况有了更宏观和全景式的把握，并精确发现：域外汉籍研究在近几年已形成一门有核心作者群和鲜明研究方向的学科，且热点与前沿方向的发展变化揭示了领域内研究工作已经出现显著的学科转向。其从“汉籍回归普查”的研究，转向为“汉籍应用转化”研究。这表明该领域已经逐渐从“引回来”研究，过渡到“用起来”和“再次走出去”的相关研究。因此，如何基于数字化手段充分展开域外汉籍开发整理、如何将域外汉籍转化为中华文化海外传播的重要材料都是应当关注的重要方向。本研究仍然存在不足，如本研究样本不包括硕博论文、专著，以及主题不符但实际研究内容为汉籍的文献，但整体来看仍能反映出学科研究的现状，为学科未来发展方向提供参考数据。

## 参考文献

Joseph C. Donohue. 2014. A bibliometric analysis of certain information science literature. *Journal of the American Society for Information Science and Technology*, 23(5):313-317.

Loet, Leydesdorff and Lutz, Bornmann ,et al. 2016. Citations: Indicators of Quality? The Impact Fallacy. *Frontiers in Research Metrics and Analytics*.

白雪涵. 2022. 基于深度学习的语义理解算法及应用研究. 北京邮电大学.

蔡艺and 谭飞and 唐群. 2022. “书籍之路”与东亚武籍交流. 体育学刊, 29(06), 31-36.

陈灏and 萧海川and 袁敏. 2023. 一个工程，一群学者，一场对海外珍稀汉籍的寻觅. 新华社, http://www.news.cn/local/2023-07/11/c_1129744215.htm.

丁学东. 1993. 文献计量学基础. 北京大学出版社.

顿泽栋. 2022. “汉籍合璧”工程中古籍印章的识别与研究. 山东大学.

傅荣贤and 华建铭. 2020. 新中国70年(1949-2019)中国古典文献学研究综述. 图书馆杂志, (02), 4-33.

胡一尘. 2020. 基于Web of Science大规模文献数据的高引论文的影响因素研究. 西南大学.

奎欣怡. 2023. 2022年中国域外汉籍研究述评. 中国文学年鉴.


(c) 2024 中国中文信息学会计算语言学专业委员会

金程宇. 2010. 近十年中国域外汉籍研究述评. 南京大学学报：哲学．人文科学．社会科学, 47(3):14.

李杰and 陈超美. 2016. CiteSpace:科技文本挖掘及可视化:text mining and visualization in scientific literature. 首都经济贸易大学出版社.

李琳. 2014. 新时期以来中国古代文学的文献研究. 复旦学报*(*社会科学版*)*, (02), 57-64.

刘蕊and 岑咏芳. 2019. 伯希和1933年北平购书与法兰西学院汉学研究所的藏书. 国际汉学, (03), 152-159, 207-208.

刘泰廷. 2017. 中国近五年域外汉籍研究述评. 图书馆理论与实践, (01), 67-72.

罗渝and 黄璜. 2018. 中文古籍数字化建设的实践与思考——以“域外汉籍数字服务平台”为例. 出版发行研究, (07), 39-42.

潘超. 2023. 中日汉籍出版中的“鳌头本”. 古代文明, 17(02), 141-155, 160.

邱均平and 杨思洛and 宋艳辉. 2012. 知识交流研究现状可视化分析. 中国图书馆学报,38(02), 78-89.

王子成. 2023. 文化传播视域下《世说新语》的经典化. 明清小说研究, (02), 180-195.

谢辉. 2015. 英国图书馆所编汉籍目录初探. 新世纪图书馆, (02), 75-79.

谢辉. 2022. 梵蒂冈图书馆收藏中国传统古籍的历程与价值. 图书馆论坛, 42(10), 150-160.

谢辉. 2016. 欧洲图书馆所编早期汉籍目录初探. 图书馆理论与实践, (02), 96-100.

严润开. 2001. 美国著名的华文汉籍文献收藏的特殊文库. 情报杂志, (04), 99-100.

杨海峥. 2022. 顾盛与北美汉籍收藏. 图书馆论坛, 42(01), 131-136.

姚委委and 金长善and 程钰情. 2021. 新媒体时代中韩文化交流与传播路径探析——推荐《域外汉籍传播与中韩词学交流》. 新闻记者, (10), 97.

张伯伟. 2006. 域外汉籍研究:域外汉籍研究——一个崭新的学术领域. 学习与探索,(02), 159-160.

张敏and 董强. 2011. 数字时代海外中华古籍的回归. 图书馆建设, (09), 38-40, 44.

郑诚. 2023. 俄罗斯国立图书馆藏中文古地图研究. 自然科学史研究, 42(02), 170-190.

钟文娟. 2012. 基于普赖斯定律与综合指数法的核心作者测评——以《图书馆建设》为例. 科技管理研究, 32(02), 57-60.

周明初and 彭志. 2014. 日本所藏稀见明词续补. 南京师范大学文学院学报, (02), 173-179.

朱本军. 2020. 海外汉籍数字化加工现状与实践研究. 古籍保护研究（01）18.

## 附录A.域外汉籍研究高被引文献

## 附录B.核心作者综合指数表


(c) 2024 中国中文信息学会计算语言学专业委员会

| 序号 | 文章标题 | 关键词 | 发表时间 | 被引量 |
|---|---|---|---|---|
| 1 | 域外汉籍与中国文学研究 | 原文无 | 2003 | 39 |
| 2 | 中日韩越古医籍数据的比较研究 | 汉字文化；古医籍；定量分析；比较研究；医学本国化；共同现象 | 2010 | 29 |
| 3 | 从“汉籍”到“域外汉籍” | 书籍之路；和制汉语；和而不同；和刻本；“域外汉籍” | 2011 | 28 |
| 4 | 近十年中国域外汉籍研究述评 | 原文无 | 2010 | 27 |
| 5 | 中国古代蒙学典籍海外传播和影响研究 | 蒙学典籍；传播；海外 | 2019 | 26 |
| 6 | 域外汉籍研究——一个崭新的学术领域 | 域外汉籍；东亚汉文小说；东亚汉诗；中、日、韩民族戏剧；拟效；汉籍东传、域外访书 | 2006 | 24 |
| 7 | 早期的汉外字典——梵蒂冈馆藏西士语文手稿十四种略述 | 汉外字典；西方汉语研究史；梵蒂冈馆藏汉籍 | 2007 | 21 |
| 8 | 日藏稀见汉籍《中兴禅林风月集》及其文献价值 | 原文无 | 2004 | 20 |
| 9 | 越南汉籍《往津日记》及其史料价值评介 | 《往津日记》；阮述；中越关系；轮船招商局 | 2010 | 20 |
| 10 | “燕行录”戏曲史料的学术价值初探 | 燕行录；戏曲史料；域外汉籍 | 2013 | 20 |

| 序号 | 作者 | 发文量 | 平均发文量 | 被引量 | 平均被引量 | 综合指数 |
|---|---|---|---|---|---|---|
| 1 | 张伯伟 | 14 | 4.68 | 99 | 17.5 | 432.4298 |
| 2 | 巴兆祥 | 4 | 4.68 | 51 | 17.5 | 188.4493 |
| 3 | 卞东波 | 12 | 4.68 | 20 | 17.5 | 185.348 |
| 4 | 金程宇 | 5 | 4.68 | 44 | 17.5 | 179.1331 |
| 5 | 陈翀 | 12 | 4.68 | 15 | 17.5 | 171.0623 |
| 6 | 刘玉 | 6 | 4.68 | 28 | 17.5 | 144.1026 |
| 7 | 王志强 | 3 | 4.68 | 37 | 17.5 | 137.7656 |
| 8 | 毛建军 | 3 | 4.68 | 33 | 17.5 | 126.337 |
| 9 | 任晓霏 | 3 | 4.68 | 32 | 17.5 | 123.4799 |
| 10 | 叶少飞 | 9 | 4.68 | 9 | 17.5 | 121.8681 |
| 11 | 王晓平 | 3 | 4.68 | 29 | 17.5 | 114.9084 |
| 12 | 钱婉约 | 4 | 4.68 | 25 | 17.5 | 114.1636 |
| 13 | 陈恒新 | 4 | 4.68 | 24 | 17.5 | 111.3065 |
| 14 | 左江 | 6 | 4.68 | 12 | 17.5 | 98.38828 |
| 15 | 宋莉华 | 3 | 4.68 | 22 | 17.5 | 94.90842 |
| 16 | 孙卫国 | 4 | 4.68 | 18 | 17.5 | 94.16361 |
| 17 | 宋丽娟 | 3 | 4.68 | 21 | 17.5 | 92.05128 |
| 18 | 严明 | 3 | 4.68 | 20 | 17.5 | 89.19414 |
| 19 | 河野贵美子 | 5 | 4.68 | 10 | 17.5 | 81.99023 |
| 20 | 陈东辉 | 4 | 4.68 | 12 | 17.5 | 77.02076 |
| 21 | 水上雅晴 | 3 | 4.68 | 12 | 17.5 | 66.337 |
| 22 | 国威 | 6 | 4.68 | 0 | 17.5 | 64.10256 |
| 23 | 叶杨曦 | 5 | 4.68 | 1 | 17.5 | 56.27595 |
| 24 | 定源(王招国) | 4 | 4.68 | 4 | 17.5 | 54.16361 |
| 25 | 顾飞 | 3 | 4.68 | 7 | 17.5 | 52.05128 |
| 26 | 付星星 | 4 | 4.68 | 0 | 17.5 | 42.73504 |
| 27 | 羽离子 | 3 | 4.68 | 3 | 17.5 | 40.62271 |
| 28 | 俞士玲 | 3 | 4.68 | 3 | 17.5 | 40.62271 |
| 29 | 杨洋 | 3 | 4.68 | 2 | 17.5 | 37.76557 |
| 30 | 何仟年 | 3 | 4.68 | 2 | 17.5 | 37.76557 |
| 31 | 童岭 | 3 | 4.68 | 1 | 17.5 | 34.90842 |
| 32 | 吴伟明 | 3 | 4.68 | 0 | 17.5 | 32.05128 |
| 33 | 刘芳亮 | 3 | 4.68 | 0 | 17.5 | 32.05128 |
| 34 | 刘一 | 3 | 4.68 | 0 | 17.5 | 32.05128 |

CCL 2024

# 面向中文多方对话的机器阅读理解研究

**蒋玉茹[1], 李宇[1*], 那婷婷[1], 张仰森[1]**
1.北京信息科技大学计算机学院，北京
{jiangyuru,2021020666}@bistu.edu.cn

## 摘要

在机器阅读理解领域，处理和分析多方对话一直是一项具有挑战性的研究任务。鉴于中文语境下相关数据资源的缺乏，本研究构建了DialogueMRC数据集，旨在促进该领域的研究进展。DialogueMRC数据集作为首个面向中文多方对话的机器阅读理解数据集，包含705个多方对话实例，涵盖24451个话语单元以及8305个问答对。区别于以往的MRC数据集，DialogueMRC数据集强调深入理解动态的对话过程，对模型应对多方对话中的复杂性及篇章解析能力提出了更高的要求。为应对中文多方对话机器阅读理解的挑战，本研究提出了融合篇章结构感知能力的中文多方对话问答模型（Discourse Structure-aware QA Model for Chinese Multi-party Dialogue，DSQA-CMD），该模型融合了问答和篇章解析任务，以提升对话上下文的理解能力。实验结果表明，相较于典型的基于微调的预训练语言模型，DSQA-CMD模型表现出明显优势，对比基于Longformer的方法，DSQA-CMD模型在MRC任务的F1和EM评价指标上分别提升了5.4%和10.0%；与当前主流的大型语言模型相比，本模型也展现了更佳的性能，表明了本文所提出方法的有效性。

**关键词：** 中文多方对话；机器阅读理解；篇章解析

# Research on Machine Reading Comprehension for Chinese Multi-party Dialogues

**Yuru Jiang[1], Yu Li[1*], Tingting Na[1], Yangsen Zhang[1]**
1. School of Computer Science,
Beijing Information Science and Technology University
{jiangyuru,2021020666}@bistu.edu.cn

## Abstract

In the field of machine reading comprehension (MRC), processing and analyzing multi-party dialogues remains a challenging research task. Given the lack of related data resources in the Chinese context, this study has developed the DialogueMRC dataset to facilitate progress in this area. As the first MRC dataset tailored to Chinese multi-party dialogues, the DialogueMRC dataset comprises 705 multi-party dialogue instances, encompassing 24,451 utterance units and 8,305 question-answer pairs. Unlike previous MRC datasets, the DialogueMRC dataset emphasizes a deep understanding of dynamic conversational processes, imposing higher demands on models in handling the

* 通讯作者
©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版
**基金项目**：北京市自然科学基金项目（4242019），题目：智慧养老服务中对话机器人关键技术研究

complexities of multi-party dialogues and discourse parsing capabilities. To address the challenges of Chinese multi-party dialogue MRC, this study introduces a discourse structure-aware question answering model for Chinese multi-party dialogue (DSQA-CMD). This model integrates question-answering and discourse parsing tasks to enhance the comprehension of dialogue context. Experimental results show that the DSQA-CMD model exhibits significant advantages over typical fine-tuned pretrained language models. Compared to methods based on Longformer, the DSQA-CMD model shows improvements of 5.4% and 10.0% in F1 and EM metrics, respectively. Moreover, the model outperforms current mainstream large language models, demonstrating the effectiveness of the proposed approach.

**Keywords:** Chinese Multi-party Dialogue , Machine Reading Comprehension , Discourse Parsing

## 1 引言

随着人工智能和自然语言处理技术的飞速发展，机器阅读理解（Machine Reading Comprehension，MRC）在智能信息获取、问答系统等领域中扮演着关键角色。尤其是将对话文本作为一种日益重要的信息载体，其MRC任务远比格式化良好的正式文档更加复杂且具有挑战性。对话文本MRC面临的主要挑战体现在模型需要深入理解动态的对话过程、处理信息的多维度特性以及准确识别对话意图上。在中文多方对话领域，由于其固有的交叉依赖特性，这些挑战变得尤为明显。目前，相关研究大多集中在英语资源上，而面向中文资源的研究则明显不足。鉴于此，本研究提出了DialogueMRC数据集，这是一个专为中文多方对话设计的片段抽取式MRC数据集。该数据集旨在促进MRC模型对中文多方对话文本结构的理解，以及提升问题答案提取的准确度，图1展示了DialogueMRC数据集经过标注的一个问答样例，其中不仅包含可回答的问题，还包含无法回答的问题。对于可回答的问题，每个问题的答案都对应上下文中的一段连续文本片段。此外，图中还展示了之前研究者所完成的篇章解析（Discourse Parsing，DP）标注成果(Jiang et al., 2023)。基于此，本研究提出了创新的多任务学习模型，旨在更有效地解析和处理复杂的对话结构以提升MRC模型的性能。本文研究的主要贡献如下：

1)提出了多阶段MRC数据标注流程，并借此构建了DialogueMRC数据集[0]，这是首个面向中文多方对话的大规模MRC数据集，包含8305个问答对，旨在提供一个全面且真实的中文对话理解环境。

2)提出了融合篇章结构感知能力的中文多方对话问答模型（DSQA-CMD模型），通过联合对话问答任务和篇章解析任务来增强模型对于复杂对话结构的理解能力。

3)实验结果表明，相较于典型的基于微调的预训练语言模型，DSQA-CMD模型表现出明显优势；与当前主流的大型语言模型相比，本模型也展现了更佳的性能，证明了本文所提出方法的有效性。

## 2 相关工作

### 2.1 机器阅读理解

机器阅读理解任务旨在训练机器阅读并理解文章内容，以便对给定问题从相关文本中找到答案(张开颜et al., 2021)。Hermann et al. (2015)于2015年构建了大规模数据集CNN/DailyMail，并提出了基于注意力的LSTM模型——THE ATTENTIVE READER，这标志着机器阅读理解技术进入神经网络时代。虽然CNN/DailyMail数据集的规模足以满足深度学习模型的训练需求，但由于其完形填空的问题形式并不符合人类自然语言的使用方式。为了突破这一限制，Rajpurkar et al. (2016)在2016年推出了SQuAD数据集，这是首个大规模机器阅读理解数据集，每个问题的答案都直接来自相关段落文本。而在中文机器阅读理解领域，由于数据资源的匮乏发展相对滞后。最早的重要贡献包括Cui et al. (2016)手工构建的首个中文阅读理解数据集，涵盖14个领域的121篇文章。此外，哈工大讯飞联合实验室（HPL-RC）

[0]https://github.com/LIyu810/DialogueMRC


(c) 2024 中国中文信息学会计算语言学专业委员会

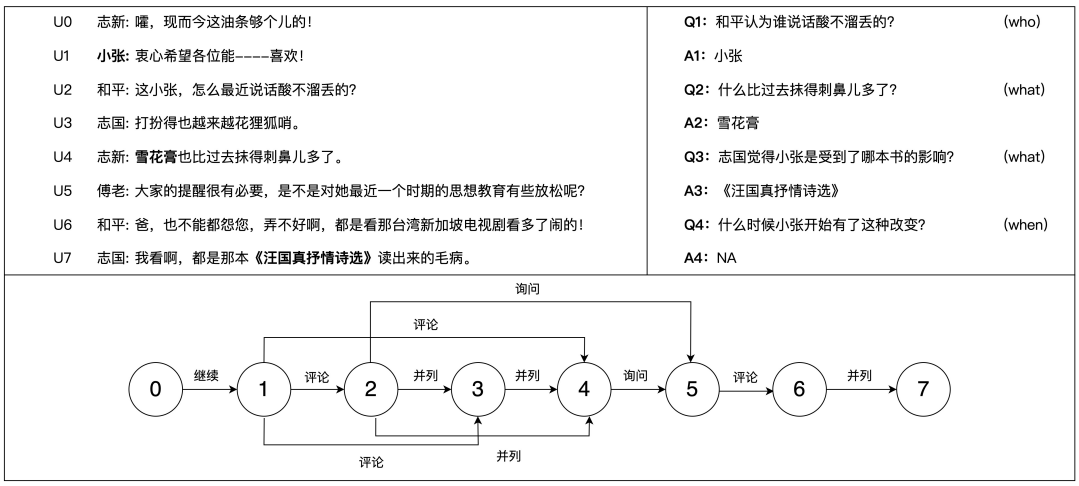



图 1. DialogueMRC数据集样例

于2019年使用人民日报和儿童读物作为语料发布了大规模中文机器阅读理解数据集，Duan et al. (2019)还基于该数据集开发了效果良好的基于注意力机制的深度阅读理解模型。

### 2.2 篇章解析

篇章解析是自然语言处理领域的基础任务，旨在分析文本中的语篇结构和语义关系，这一任务涵盖了诸如评论、致谢及问答等多种关系类型。目前，篇章解析技术主要基于宾州树库（PDTB）和修辞结构理论树库（RST-DT）等数据集进行研究。PDTB(Webber et al., 2019)和RST-DT(Hou et al., 2020)的发布显著推动了语篇关系及篇章语义分析技术的发展，这两个语料库的原始数据均来源于华尔街日报。然而，随着研究的深入，学界对于偏口语化的多方对话篇章解析表现出增加的兴趣。在这一领域，STAC数据集(Afantenos et al., 2015)成为主流语料库之一。该数据集由Afantenos等人在2016年发布，基于游戏《The Settlers of Catan》的在线版本进行构建。此游戏涉及多玩家的合作与竞争交流，为篇章分析提供了丰富的口语化文本资源。

### 2.3 融合篇章解析的机器阅读理解研究

近年来，将篇章解析融入机器阅读理解的研究受到了广泛关注。Molweni数据集(Li et al., 2020)作为该研究领域的重要里程碑，包含10000个多方对话，88303个篇章单元和30066个问答对，此外，研究者还标注了78245个篇章关系。Molweni要求模型不仅需要理解每个发言的表面文字信息，还要分析不同篇章单元之间的篇章关系，从而实现对上下文更深入的理解。Molweni数据集来源于Ubuntu操作系统的在线技术论坛，其相较于日常生活场景的发言内容，在交流形式上存在显著差异。通常，日常对话中的发言内容更加普遍，一次只能由一人发言，而且对话是即时进行的，不具有滞后性。这些特征赋予了基于日常场景的多方对话数据集其独特的价值，为了更好的挖掘这些日常生活中对话的特点，Jiang et al. (2023)提出了DialogueDSA的数据集。该数据集精心挑选并标注了包括因果、转折、并列等在内的16种篇章关系类型。相关研究表明，篇章结构的理解对于机器准确回答问题至关重要，特别是在需要长距离推理和把握复杂对话流的情境下。此外，针对Molweni的实验结果也揭示了DP具有提高MRC任务性能的潜力(He et al., 2021)。在本文中，将进一步探讨融合篇章解析的机器阅读理解技术，以及该技术在处理多方对话文本中的应用效果和潜在的改进方向。

## 3 DialogueMRC数据集的构建与分析

本文所构建的数据集（以下简称为DialogueMRC）语料来自1994年上映的120集情景喜剧《我爱我家》的剧本文本。该剧情景包含117位出场角色，其中8位为主要角色。下述列出了选择《我爱我家》剧本文本作为标注基础语料库的三个原因。

1)该剧内容与日常生活紧密相连，涵盖了广泛的话题，如家庭关系、职场矛盾、邻里交往等。这种丰富多彩的生活场景为机器阅读理解研究提供了一个接地气的、现实生活中的应用背


(c) 2024 中国中文信息学会计算语言学专业委员会

景，非常适合探索和解决日常生活场景下的机器阅读理解问题。

2)该剧本涵盖了117位角色，每个角色都有其独特的性格、语言风格和情感表达。这种角色多样性对数据集的标注提出了挑战，同时也提供了丰富的标注维度，这对于构建一个能够理解复杂人物关系和情感交互的机器阅读理解模型至关重要。

3)目前已有研究基于《我爱我家》的剧本文本标注了关系抽取(Jiang et al., 2022)、篇章解析(Jiang et al., 2023)等数据集，并在此基础上完成了相关实验，提出了CSE-DPM(Jiang et al., 2023)等模型，这些工作为本文的研究提供了坚实的基础。

在问答设计中，DialogueMRC特别考虑到多方对话的复杂性，其中对话参与者可能多于两人。这一设定增加了问题设计的难度，因为需要准确识别和理解每个参与者的发言及其在对话中的角色和立场。此外，多方对话的动态性要求数据标注过程中必须考虑对话流的连贯性和上下文变化，这对于提高模型在真实场景中的应用能力至关重要。通过这种针对性的设计，DialogueMRC旨在提升模型对复杂对话结构的理解力，从而更好地服务于多参与者交互的实际应用场景。

DialogueMRC数据集的场景选择继承自CRECIL和DialogueDSA所采用的同一批705个对话场景，这种继承性不仅保证了数据的一致性，也为后续的多任务联合训练研究提供了便利。其中每个场景都被视作一个独立的对话单元，以便于进行更精确的数据处理与分析。DialogueMRC可以被看作是片段抽取任务，即在给定的对话上下文中提出问题，并要求模型在对话内容中找到答案片段。为了确保DialogueMRC数据集在精确性和可靠性方面达到最高标准，本研究提出了多阶段机器阅读理解数据标注流程，并建立了详尽的数据质量检查体系。通过这些措施，力求使DialogueMRC数据集成为机器阅读理解领域的一个可靠和有价值的资源。

### 3.1 多阶段机器阅读理解数据标注流程

在数据标注领域，人工标注因其提供高度准确且可靠的数据而长期被视为标准。但是，人工标注的缺点是成本高且耗时。随着人工智能技术的发展，使用文本生成技术进行数据标注成为了一种新兴方法，利用类似GPT-3.5的大语言模型可以高效地生成大量标注数据，提高了标注效率。但这种方法的挑战在于生成的标签可能包含噪声，特别是在特定领域或复杂任务中更为明显。鉴于以上挑战，本文提出了一种多阶段机器阅读理解数据标注流程，如图2所示，旨在结合人工标注的准确性与使用大语言模型自动标注的高效性，以达到更优的标注结果。

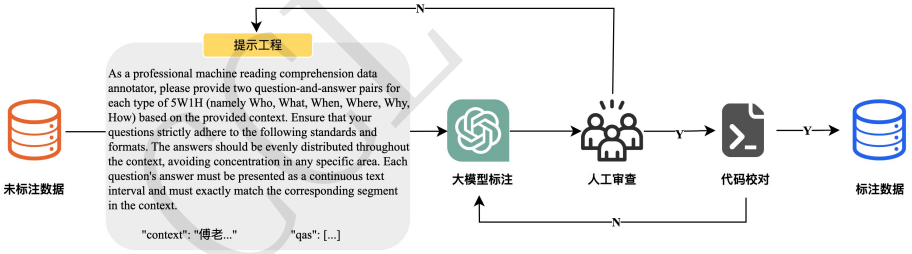



图 2. 多阶段机器阅读理解数据标注流程图

1)提示工程：流程的首要步骤是设计初始的提示。此步骤涉及输入特定的提示和对话内容，引导大型语言模型生成相应的问答对。然后，通过人工审查这些问答对，识别并解决诸如问题类型不匹配、问答对在对话中分布不均、生成的答案与上下文连续片段不匹配等问题。通过这种持续的反馈和优化过程，逐步完善提示语的设计，确立数据标注的准确提示语。

2)大语言模型标注：本研究选择目前业界前沿的大语言模型ChatGPT-4.0来完成标注工作，此举旨在提升标注效率(Gilardi et al., 2023)。借助大语言模型强大的语言理解能力，本研究高效地标注了大量的高质量问答对。该模型通过分析对话内容生成不同类型的问题，然后根据上下文内容生成准确的答案。这一步骤大幅减少了传统手工标注所需的时间和人力成本。

3)人工审查：鉴于大型语言模型可能出现的幻觉等问题，为确保标注数据的高质量，本研究采用人工审查方式。审查内容包括模型生成问答对的类型匹配、5W1H（何时、何地、何人、何物、为何、如何）问题的数量分布、有答案与无答案问题的准确性和鲁棒性等。在发现问答对存在问题时，标注人员将对提示语进行微调，并重新生成问答对，直至通过人工审查。


(c) 2024 中国中文信息学会计算语言学专业委员会

4)代码校对：通过编写专门的脚本，对生成的答案进行审核，确保答案是对话上下文中的连续区间，并检查答案的起始位置是否准确。若发现错误，则对错误样例重新进行大语言模型标注，直至通过代码校对。

总结而言，本文提出的多阶段机器阅读理解数据标注流程，有效地融合了手工标注的精准性与自动化标注的高效性。

### 3.2 DialogueMRC语料库数据分析

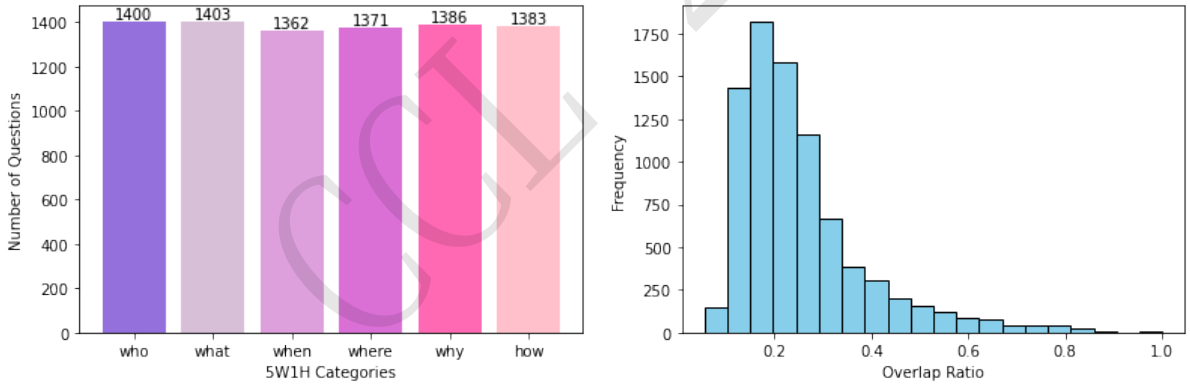
标注工作完成后，得到了一个详尽且均衡的机器阅读理解数据集，如表1所示，DialogueMRC数据集包含了总计705个对话实例，其中训练集占565个，开发集和测试集各占70个。

| | 训练集 | 验证集 | 测试集 |
|---|---|---|---|
| 对话数量 | 565 | 70 | 70 |
| 话语单元数量 | 19556 | 2576 | 2319 |
| 问题数量 | 6654 | 830 | 818 |
| 可回答问题数量 | 5526 | 662 | 689 |
| 不可回答问题数量 | 1128 | 168 | 129 |

表 1. DiologueMRC数据集概览

#### 3.2.1 5W1H问题类型分布

DialogueMRC数据集中问题类型在5W1H范式下的分布均衡，如图3(a)所示。这种均衡对于确保机器阅读理解系统的全面性和鲁棒性至关重要，有助于系统对各类问题进行平等处理，减少偏差。此外，数据集的精心设计与平衡标注有利于发展广泛适用的模型，并公平评估各类问题上的模型性能。




(a) 5W1H问题类型分布图　　(b) 问题与上下文的词汇重叠率分布图

图 3. 问题类型及词汇重叠分析示意图

#### 3.2.2 问题与上下文的词汇重叠率分布

图3(b)展示了DialogueMRC数据集中问题与其上下文之间词汇重叠率的分布情况。词汇重叠率是衡量问题词汇与其上下文共享程度的指标，这一指标反映了问题相对于对话内容的复杂性和独特性。在该数据集中，一个明显的高频区间发生在10%到30%的词汇重叠率范围内，表明相当一部分问题与其上下文共享了适度的词汇量。在更高的重叠率范围内，频率显著下降，只有2个问题的重叠率在90%至100%之间。这一急剧的下降表明与上下文几乎完全相同的问题极为罕见，这意味着大多数问题需要对对话内容有更深层次的理解才能确定正确答案。该问题分布证明了DialogueMRC语料库经过了精心设计，提供了有效训练机器阅读理解模型的数据。

第二十三届中国计算语言学大会论文集，第650页-第661页，太原，中国，2024年7月25日至28日。
卷1：主会论文
(c) 2024 中国中文信息学会计算语言学专业委员会

## 4 融合篇章结构感知的多方对话问答模型设计

本研究提出了融合篇章结构感知能力的中文多方对话问答模型（Discourse Structure-aware QA Model for Chinese Multi-party Dialogue，DSQA-CMD），如图4所示，问答任务通过预训练模型的输出来预测答案在文本中的位置，而篇章解析任务则通过构建依赖树捕获话语间的结构关系，这两个任务的联合训练旨在加深对多方对话的理解，从而提升MRC模型的性能。

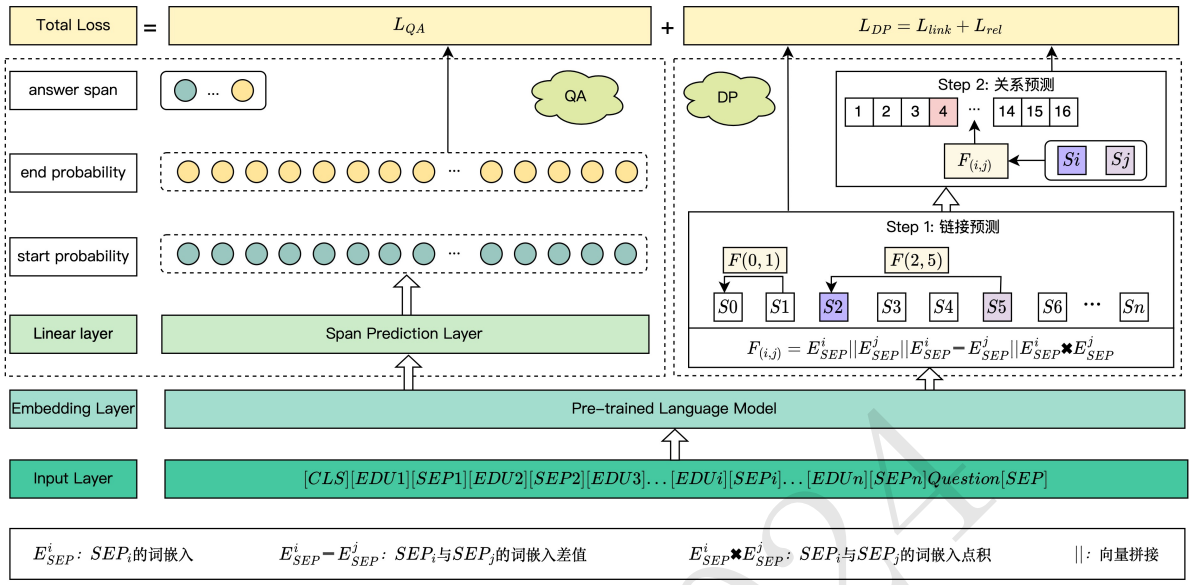



图 4. 融合篇章结构感知能力的中文多方对话问答模型图

### 4.1 输入特征表示

对话上下文与问题在预处理阶段通过预训练语言模型进行编码。为明确区分问题与对话内容以及对话中的不同篇章单元，特殊分隔符号（$[SEP]$）被插入相应位置。这样不仅有助于辨识问题与上下文的边界，还为篇章解析任务提供了话语链接预测的关键线索。同时，对话中的每个分隔符位置都被准确记录，以确保篇章解析任务中能够单独提取出每个话语对应的特征表示。最后，问题Q与对话中所有的篇章单元（$EDU$）连同特殊符号一同作为输入，得到融合了问题与上下文信息的序列特征表示S，具体表示为$S = encoder\left([CLS], [EDU1], [SEP1], \ldots, [EDUn], [SEPn], Q, [SEP]\right)$

### 4.2 问答任务

问答任务直接使用从预训练语言模型中得到的输出特征。模型通过全连接层为给定对话中的每个词元（token）预测作为答案开始和结束的概率，并据此从对话文本中提取答案片段。起始位置的预测如公式1所示，其中$P_s$表示预测的起始位置，下标$i$表示词元序列中的索引，$W_s$为用于预测起始位置的权重矩阵，$S$为从预训练语言模型中获得的文本特征。在处理无法回答的问题时，首先计算最可能的答案范围分数，并将其与无答案的预设阈值进行比较，以此来决定问题是否有确定答案。

$$P_s = argmax_i \left(softmax\left(W_s S\right)_i\right) \tag{1}$$

问答任务旨在准确预测答案在文本中的起始和结束位置。该任务被构建为一个分类问题，假定输入嵌入中共有$M$个词元，每个词元对应一个可能的起始或结束概率。利用$softmax$函数对模型输出的logits进行归一化，得到各个位置的概率分布。然后通过交叉熵损失函数计算起始和结束位置的预测损失。如公式2所示，$M$表示问答任务的标签数量，起始位置标签$y_s^m$为1表示第$m$个词元为开始位置，否则为0。同样，结束位置标签$y_e^m$为1表示第$m$个词元为结束位置，否则为0。模型预测的起始位置概率为$P_s^m$，结束位置概率为$P_e^m$。QA任务的总损失$L_{QA}$通过起始

和结束位置预测损失的均值来表示。

$$L_{QA} = -\frac{1}{2}\sum_{m=0}^{M-1}(y_s^m logP_s^m + y_e^m logP_e^m) \tag{2}$$

### 4.3 篇章解析任务

篇章解析任务需要进一步处理经过预训练语言模型得到的特征，以表征话语之间的依赖关系。特别地，篇章单元所对应的分隔符向量被选定为话语特征的代表，进一步计算出用于篇章解析的关系特征向量。这一过程结合了欧几里得距离和余弦相似度的计量，旨在更全面地捕捉话语间的细微关系。为了解析篇章单元$[EDUi]$和$[EDUj]$所对应的篇章关系，本模型构建了特征向量的联合表示$F_{i,j}$（$F_{i,j} = \left(E_{SEP}^i, E_{SEP}^j, E_{SEP}^i - E_{SEP}^j, E_{SEP}^i \times E_{SEP}^j\right)$）。其中，$E_{SEP}^i$和$E_{SEP}^j$分别对应篇章单元$[EDUi]$和$[EDUj]$的输出特征，$F_{i,j}$代表对不同表示的拼接，模型综合了分隔符对应的向量及其交互信息，从而更好地揭示篇章单元之间的深层结构关系。

篇章解析任务通过构建依赖树来表征话语之间的依赖关系。篇章解析任务分为两个部分：链接预测和关系预测。对于链接预测，模型旨在确定话语间是否存在依赖关系。对于每个话语，通过计算特征向量来预测其依赖对话中的哪个话语。当话语不依赖于其他话语时，将其指定为依赖于一个虚构的根节点。如公式3所示，$L_i$表示预测的依赖话语索引，$a$表示遍历所有可能的依赖目标话语的索引，$W_l$表示用于链接预测的权重矩阵。而在关系预测中，模型需确定话语间的确切关系类型，如公式4所示，$R_{i,j}$表示预测的第$i$个话语和第$j$个话语之间的关系类型，$k$表示所有可能的关系类型，$W_r$表示用于关系预测的权重矩阵。

$$L_i = argmax_a\left(softmax\left(W_l\left[F_{i,1}, F_{i,2}, \ldots, F_{i,t}\right]\right)_a\right) \tag{3}$$

$$R_{i,j} = argmax_k\left(softmax\left(W_r F_{i,j}\right)_k\right) \tag{4}$$

DialogueMRC数据集单一对话中最大基本篇章单元数量为$T\,(T=90)$，篇章关系类型数量为16。链接预测和关系预测被视为多类分类问题，分别具有$T+1$和16个类别，其中$T+1$中的额外一个标签代表根节点。交叉熵损失函数用于分别计算链接预测$L_{link}$和关系预测$L_{rel}$的损失，对其求和得到DP任务的总损失$L_{DP}$。

链接预测的损失函数定义如公式5所示，如果第$t1$个话语依赖于第$t2$个话语，那么$y_l^{t1,t2}$为1，否则为0。$p_l^{t1,t2}$表示模型预测第$t1$个话语依赖于第$t2$个话语的概率值。

$$L_{link} = -\frac{1}{T}\left(\sum_{t1=0}^{T-1}\sum_{t2=0,t2\neq t1}^{T-1} y_l^{t1,t2} logP_l^{t1,t2}\right) \tag{5}$$

关系预测的损失函数$L_{rel}$如公式6所示，其中，如果第$n$个样本中第$t$个话语与任何其他话语存在第$i$种类型的关系，则$y_r^{t,i}$等于1，否则等于0。同时，$p_r^{t,i}$表示模型预测第$t$个话语与任何其他话语存在第$i$种关系的概率。

$$L_{rel} = -\frac{1}{T}\left(\sum_{t=0}^{T-1}\sum_{i=0}^{15} y_r^{t,i} logP_r^{t,i}\right) \tag{6}$$

如公式7所示，DP任务的总损失$L_{DP}$是链接预测和关系预测损失的和。

$$L_{DP} = L_{link} + L_{rel} \tag{7}$$

将DP任务的总损失$L_{DP}$与QA任务的损失$L_{QA}$相加，得到DSQA-CMD模型的总损失$L$。

$$L = L_{QA} + L_{DP} \tag{8}$$


(c) 2024 中国中文信息学会计算语言学专业委员会

# 5 实验与分析

## 5.1 评价指标

实验采用F1分数和精确匹配（EM）来作为问答（QA）任务的评价指标。F1分数是基于精确率（P）和召回率（R）的调和平均值，公式定义如9所示。其中，精确率度量了预测为正例的样本中真实为正的比例，召回率度量了在所有真实正例中被正确预测为正的比例。

$$F1 = \frac{2 \times P \times R}{P + R} \tag{9}$$

EM（精确匹配）是在问答系统中广泛应用的评价指标，用于衡量预测答案与任何一个真实答案完全一致的比例，公式定义如公式10所示。对于篇章解析任务，本研究使用F1分数来分别评估链接预测和关系预测的性能。在关系预测中，只有当链接和关系预测都正确时，才会被视为结果正确。

$$EM = \frac{\text{完全匹配的预测数量}}{\text{总预测数量}} \tag{10}$$

## 5.2 实验设置

本研究选取两种预训练语言模型，分别为RoBERTa(Liu et al., 2019)和MacBERT(Cui et al., 2020)，作为对比实验的基础。这两种模型分别被设定为实验的基线（benchmark）和当前最优（state-of-the-art）性能指标，这两种模型均具有768维的隐藏层。在对话上下文处理中，由于数据集存在上下文长度超过512个词元的实例，为适应模型输入限制并保持上下文完整性，本研究设置最大序列长度为512，并采用文档滑动窗口为128的策略。此外，问题和答案的最大长度分别设定为64和60，以涵盖大多数潜在答案。实验中设置每个对话的最大话语数为40，模型训练采用2e-5的学习率，并设定丢弃率为0.2，以抑制过拟合现象。权重衰减设置为0.01，以进一步正则化模型训练。经过初步实验验证，所有模型均在10轮内完成训练，批次大小为18。

## 5.3 对比模型

### 5.3.1 基于微调的预训练对比模型

为了全面评价本研究提出的多任务学习模型在多方对话机器阅读理解任务上的表现，本研究设计了一系列对比实验，涉及多种基于微调的深度学习模型。本研究基于MacBERT构建了多任务学习模型（以下简称DSQA-CMD（Mac））和一个采用RoBERTa作为预训练模型的多任务模型（以下简称DSQA-CMD（Ro））。此外，为了确保评估的全面性，本研究还同其他专门针对问答和篇章解析任务的模型作为对比，包括针对长文本的Longformer机器阅读理解模型，以及专注于篇章解析的Deep sequential模型和CSE-DPM模型。这些对比实验不仅有助于揭示DSQA-CMD (Mac)在多方对话MRC任务中的相对优势，也为深入理解不同模型架构和训练策略在复杂对话环境中的适应性和效果提供了丰富的数据支持。

### 5.3.2 基于提示驱动的大语言对比模型

在自然语言处理领域，大型预训练语言模型如GPT-3.5-turbo已广泛应用于多种任务，并展现出卓越的性能。本研究将进一步探索这些模型在MRC任务中的应用潜力，特别是在处理DialogueMRC数据集时的表现。本研究采用基于提示的评测方法，该方法相较于传统微调，允许模型直接在多样化对话场景中进行理解和推理，无需进行特定的微调。为此，本研究设计了一系列提示语，引导模型准确理解并回答源自复杂多方对话的问题。

为了全面评估方法的有效性，本研究选取了几种最新的大型模型进行对比，包括ChatGLM3-6b(Du et al., 2022)、Vicuna-13b(Zheng et al., 2024)、Baichuan2-7b(Yang et al., 2023)、GPT-3.5-turbo(Brown et al., 2020)以及Qwen1.5-72b-chat(Bai et al., 2023)。这些模型均通过不同的架构和预训练策略，在NLP任务中取得了显著成绩。通过与上述模型的综合比较，本研究旨在探索基于提示的评测方法在激发模型深层次理解和推理能力方面的有效性。

## 5.4 实验结果与分析

### 5.4.1 与基于微调的预训练模型的对比结果与分析

如表2所示，在QA任务中，DSQA-CMD(Mac)模型分别在F1和EM指标上达到了45.9%和36.3%，显著优于先前的模型。特别是相较于使用传统方法，如不对文本进行切分处


(c) 2024 中国中文信息学会计算语言学专业委员会

理，同采用基于Longformer(Beltagy et al., 2020)架构的方法相比，DSQA-CMD(Mac)在F1指标上提高了5.4%，EM指标上提升了10.0%。这一结果突显了在问答任务中，多任务训练模型在性能均衡和优势方面的显著优越性。

| 模型 | QA | | DP | |
|---|---|---|---|---|
| | F1(%) | EM(%) | Link(%) | Link&Rel(%) |
| Longformer | 40.5 | 26.3 | - | - |
| Deep sequential | - | - | 73.1 | 42.2 |
| CSE-DPM | - | - | **76.2** | **46.7** |
| DSQA-CMD(Ro) | 41.1 | 33.0 | 61.4 | 36.1 |
| DSQA-CMD(Mac) | **45.9** | **36.3** | 61.6 | 37.2 |

表 2. 基于微调的预训练对比模型实验结果

### 5.4.2 与基于提示驱动的大语言模型对比实验结果与分析

表3展示了多任务学习模型与当前主流大型语言模型在DialogueMRC上的性能对比。结果表明，DSQA-CMD(Mac)在F1和EM指标上均优于参考的大型模型，尤其是在EM指标上相比于最接近的Qwen1.5-72b-chat模型，实现了12.8%的显著提高，同时在F1分数上也取得了2.3%的领先。分析大型语言模型在DialogueMRC数据集上的表现时，可以发现它们在生成答案时倾向于生成完整的句子或短语。这种生成式的回答格式常常无法严格对应到原文的确切位置，导致即使答案内容正确，也因格式不符合抽取式任务的要求而被评估为错误。这一现象强调了大模型在处理具体、特定领域任务时可能的局限性。随着模型参数数量的增加，其性能也随之提升，这强调了大语言模型参数量对于增强模型理解能力和处理复杂任务方面的关键作用。此外，实验表明尽管大模型在规模上的优势有助于提升通用任务的处理能力，但当面对具体的、特定领域的任务时，针对具体下游任务设计模型和训练显得更为关键。尤其是在中文多方对话理解任务中，大语言模型可能无法充分捕捉到任务的特定需求和细微差异。而多任务学习模型通过对问答和篇章解析任务的联合优化，不仅提升了模型对对话上下文的深层理解，还增强了针对具体任务需求的适应能力，从而实现了在关键性能指标上的显著提升。

| 模型 | QA | |
|---|---|---|
| | F1(%) | EM(%) |
| ChatGLM3-6b | 16.7 | 2.9 |
| Vicuna-13b | 21.1 | 5.9 |
| Baichuan2-7b | 25.1 | 8.7 |
| GPT-3.5-turbo | 34.5 | 15.2 |
| Qwen1.5-72b-chat | 43.6 | 23.3 |
| DSQA-CMD(Mac) | **45.9** | **36.3** |

表 3. 基于提示驱动的大语言对比模型实验结果

### 5.4.3 消融实验结果及分析

为了深入了解多任务学习模型在中文多方对话机器阅读理解任务中的表现，本文设计了一系列消融实验。实验结果如表4所示，其展示了DSQA-CMD模型相较于单任务模型在F1和EM指标上的提升。实验结果有力地证实了多任务学习策略的有效性。值得注意的是，多任务模型的出色表现揭示了任务间的正向互动效应。具体来说，篇章解析任务的融合促进了模型在共享特征空间中的表示学习，从而增强了问答性能。这种正向效应可归因于两个任务在语义理解和信息提取方面的内在联系。通过联合训练，模型能够在篇章解析中学习深层次语义关系和对话结构知识，进而在问答任务中更有效地定位和推理答案。


(c) 2024 中国中文信息学会计算语言学专业委员会

| 模型 | QA | | DP | |
|---|---|---|---|---|
| | F1(%) | EM(%) | Link(%) | Link&Rel(%) |
| QA-only | 43.2 | 33.7 | - | - |
| DP-only | - | - | 60.6 | 35.3 |
| DSQA-CMD(Mac) | **45.9** | **36.3** | **61.6** | **37.2** |

表 4. 消融实验结果

### 5.4.4 基于5W1H问题类型的模型性能评估

为深入理解本文提出的多任务学习模型在处理不同类型问题时的性能，本文分析了模型在5W1H问题类型上的表现。表5展示了各问题类型的评估结果，以及它们相较于全量数据集训练的平均性能的变化。本研究发现模型在处理“Who”和“Where”问题时表现尤为出色，这主要归因于其强大的实体识别能力，特别是在定位对话中的人物和地点方面，这些问题类型通常不需要复杂推理即可有效解答。相反，“Why”和“How”问题的表现较差，主要因为这些问题需要深度理解和推理，包括因果关系和方法论的把握，此外，这类问题的答案往往较长，包含多个信息点，要求模型不仅准确理解问题意图，还需从广泛的上下文中提取准确的信息片段，这大大增加了答案精准匹配的难度。模型在这些问题类型上的局限性揭示了其在处理复杂语义、逻辑推理以及长篇信息方面的不足。

| 问题类型 | 样本数 | F1(%) | EM(%) |
|---|---|---|---|
| What | 139 | 47.4(↑1.5) | 29.5(↓6.8) |
| Who | 138 | 69.7(↑23.8) | 68.1(↑31.8) |
| Why | 138 | 27.8(↓18.1) | 12.3(↓24.0) |
| How | 134 | 24.8(↓21.1) | 9.7(↓26.6) |
| When | 134 | 51.7(↑5.8) | 47.8(↑11.5) |
| Where | 135 | 53.6(↑7.7) | 50.4(↑14.1) |

表 5. 基于5W1H问题类型的模型性能评估结果

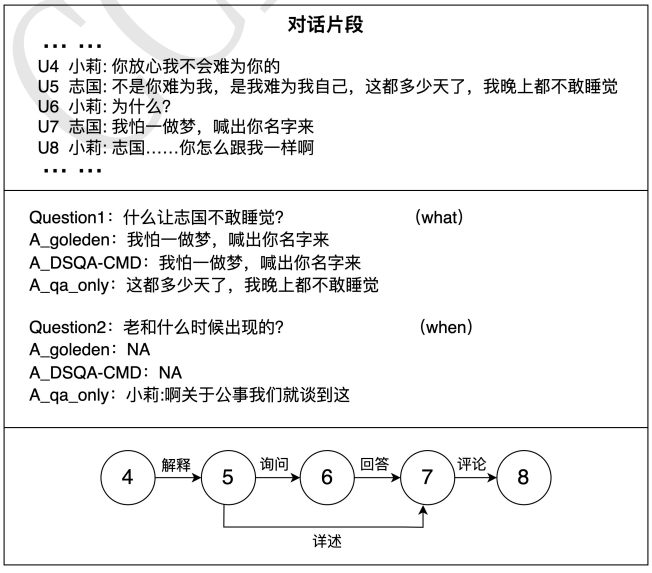



图 5. 案例分析


(c) 2024 中国中文信息学会计算语言学专业委员会

### 5.4.5 案例分析

图5所示对话片段涉及两个角色，分别是小莉和志国，他们之间的对话展示了角色之间的情感动机和个人困扰。本研究通过两个具体问题来评估模型的性能：一是询问导致志国失眠的原因，二是探究一个特定事件的出现时间。在回答“什么让志国不敢睡觉？”时，DSQA-CMD模型通过捕捉对话中的关键信息，志国担心梦中无意间喊出小莉的名字，成功地给出了问题的答案，这体现了其在理解对话中隐含的情绪和深层含义方面的出色能力。在这一对话片段中，每一轮交流都建立在前一轮的基础上，形成了逻辑上的连续性。例如，小莉的询问（U6）引出了志国的回答（U7），而志国的回答进一步详述了他的担忧（U7对U5的详述）。篇章解析的融入使模型能够识别这些细微的逻辑连接，进而在回答问题时提供更加准确和深入的理解。面对第二个问题“老和什么时候出现的？”时，由于问题存在歧义且信息不足。DSQA-CMD模型预测答案不存在，这展现了其在处理不确定性问题时的鲁棒性，同时体现了篇章解析在进一步整合和理解对话结构方面的重要作用。

## 6 结束语

本研究构建了面向中文多方对话的机器阅读理解数据集DialogueMRC，为中文MRC的研究提供了高质量的数据资源。此外，通过将篇章解析与问答任务相结合，本研究提出了融合篇章结构感知能力的中文多方对话问答模型，并在DialogueMRC数据集上证明了篇章解析在提升多方对话机器阅读理解能力方面的关键作用。本文的研究成果不仅推动了中文多方对话MRC领域的技术创新，还为智能客服、智能家居等相关应用提供了强有力的技术支持。未来，将试图融合图网络技术建模，以提出更加创新的多方对话MRC模型，为人机交互和自然语言处理的发展提供新的思路和方向。

## 参考文献

Stergos Afantenos, Eric Kow, Nicholas Asher, and Jérémy Perret. 2015. Discourse parsing for multi-party chat dialogues. pages pp–928.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yiming Cui, Ting Liu, Zhipeng Chen, Shijin Wang, and Guoping Hu. 2016. Consensus attention-based neural networks for chinese reading comprehension. *COLING 2016 - 26th International Conference on Computational Linguistics, Proceedings of COLING 2016: Technical Papers*, pages 1777 – 1786.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for chinese natural language processing. *Findings of the Association for Computational Linguistics Findings of ACL: EMNLP 2020*, pages 657 – 668.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1:320 – 335.

Xingyi Duan, Baoxin Wang, Ziyue Wang, Wentao Ma, Yiming Cui, Dayong Wu, Shijin Wang, Ting Liu, Tianxiang Huo, Zhen Hu, et al. 2019. Cjrc: A reliable human-annotated benchmark dataset for chinese judicial reading comprehension. pages 439–451.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Yuchen He, Zhuosheng Zhang, and Hai Zhao. 2021. Multi-tasking dialogue comprehension with discourse parsing. In Kaibao Hu, Jong-Bok Kim, Chengqing Zong, and Emmanuele Chersoni, editors, *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 551–561, Shanghai, China, 11. Association for Computational Linguistics.

CCL-2024

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

Shengluan Hou, Shuhan Zhang, and Chaoqun Fei. 2020. Rhetorical structure theory: A comprehensive review of theory, parsing methods and applications. *Expert Systems with Applications*, 157:113421.

Yuru Jiang, Yang Xu, Yuhang Zhan, Weikai He, Yilin Wang, Zixuan Xi, Meiyun Wang, Xinyu Li, Yu Li, and Yanchao Yu. 2022. The crecil corpus: a new dataset for extraction of relations between characters in chinese multi-party dialogues.

Yuru Jiang, Yu Li, Weikai He, Jie Chen, Yanchao Yu, and Yangsen Zhang. 2023. A new dataset and parsing model for chinese multiparty dialogue discourse structure. pages 221–227.

Jiaqi Li, Ming Liu, Min-Yen Kan, Zihao Zheng, Zekun Wang, Wenqiang Lei, Ting Liu, and Bing Qin. 2020. Molweni: A challenge multiparty dialogues-based machine reading comprehension dataset with discourse structure. *arXiv preprint arXiv:2004.05080*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, pages 2383 – 2392.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The penn discourse treebank 3.0 annotation manual. *Philadelphia, University of Pennsylvania*, 35:108.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

张开颜, 张伟男, and 刘挺. 2021. 基于深度学习的多方对话研究综述. 中国科学: 信息科学, 51(8):1217–1232.

CCL 2024

# 融合半监督学习与同义计算的传染病名称自动映射研究

**宋培彦[1] 杨青香[12*] 胡博深[1] 杜博雅[1]**
[1] 天津师范大学管理学院，天津300387
[2] 华中师范大学人工智能教育学部，武汉430079
spyer2008@126.com

## 摘要

医学古籍蕴含着丰富的专业知识，然而由于古代疾病名称、术语与现代标准表述不一致等问题，严重影响了公共卫生知识组织和服务质量，现有研究主要采用专家手工映射、词义计算等方式解决，存在着工作效率和准确率偏低等问题，以古籍术语辞典作为语料进行挖掘、建立传统医学术语与现代医学术语的同义关系，并映射到国际规范，形成“古-今-外”三语互通的知识库是可行方法。为此，本文以知识组织和知识发现理论为基础，设计了古今疾病名称跨语言自动映射方法，并以传染性疾病名称为例进行验证。具体过程是：首先，利用snowball算法抽取古今疾病名称同义模式，获取了12个与传染性疾病相关的疾病名称关系模式和134个同义词对。其次，依据桑基图从关联性、成熟度和延展性3个角度分析疾病名称历时演变进行可视化关联分析。同时，结合sapbert词向量和余弦相似度将传统医学疾病名称向ICD-11国际标准映射，经过人工验证映射结果达到0.23的hit@1、0.42的hit@5以及0.61的hit@10。本文发现，通过专业辞典语料，可以抽取疾病名称的语言变异情况，提高同义术语的发现效率，为构建专业知识库提供更多的入口词和语义关联性，缓解信息孤岛问题。研究还表明，以辞典中的现代医学术语名称作为映射起点，关联到ICD-11国际规范，为开展跨语言领域知识工程建设提供参考，对实现专业知识“古为今用“和国际传播也具有重要现实意义。

**关键词：** 关系抽取；模式匹配；术语映射；机器学习算法；术语计算

# A study on automatic mapping of infectious disease names by integrating semi-supervised learning and tautology computation

**Song Peiyan[1] Yang Qingxiang[12*] Hu Boshen[1] Du Boya[1]**
[1] School of Management,
Tianjin Normal University,Tianjin 300387
[2] Artificial Intelligence Education Department,
Central China Normal University,Wuhan430079S
spyer2008@126.com

## Abstract

Ancient medical texts contain a wealth of professional knowledge, yet the organization of public health knowledge and the quality of services are seriously affected by problems such as inconsistencies between ancient disease names and terminology and modern standard expressions. Existing studies have mainly used manual mapping and

*通讯作者
©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版
本文为2020年度国家社科基金重大项目“国家重大突发事件信息公开质量研究”（20ZD141）成果

lexical meaning computation by experts to solve the problems of low efficiency and accuracy, etc. It is feasible to use ancient terminology thesaurus as the corpus for mining, establishing synonym relations between traditional and modern medical terms, and mapping to international norms to form a trilingual knowledge base of "ancient-present-foreign". In this paper, we propose a feasible way to mine the corpus of ancient terminology, establish the synonym relationship between traditional medical terms and modern medical terms, and map them to the international norms to form a knowledge base of "ancient-modern-foreign" trilingual interoperability. In this paper, based on the theory of knowledge organisation and knowledge discovery, we design a cross-lingual automatic mapping method for ancient and modern disease names, and validate it with infectious disease names as an example. The specific process is as follows: firstly, the snowball algorithm is used to extract the synonym patterns of ancient and modern disease names, and 12 disease name relationship patterns and 134 synonym pairs related to infectious diseases are obtained. Secondly, the evolution of disease names over time was analysed from three perspectives: relevance, maturity and extensibility for visual association analysis based on Sankey diagram. Meanwhile, the traditional medical disease names were mapped to the ICD-11 international standard by combining the sapbert word vector and cosine similarity, and the mapping results reached 0.23 hit@1, 0.42 hit@5, and 0.61 hit@10 after manual validation.In this paper, it is found that the linguistic variability of disease names can be extracted from professional thesaurus corpora, and the discovery of synonyms can be improved. efficiency, provide more entry words and semantic correlations for building a professional knowledge base, and alleviate the problem of information silos. The study also shows that taking the modern medical terms in the thesaurus as the starting point for mapping and linking them to the ICD-11 international standard can provide a reference for the construction of cross-linguistic knowledge engineering, which is also of great practical significance for the realisation of professional knowledge and the international dissemination of professional knowledge.

**Keywords:** relational extraction , pattern matching , term mapping , machine learning , algorithm Terminology calculations

## 1 引言

面对重大公共卫生事件特别是传染性疾病，不同国家和地区使用了知识组织工具和术语来描述，如ICD-11、UMLS、SNOMEDCT等，为实现有效的知识服务和决策支持提供了必要的保障。为了提高数据质量、实现知识互通，2019年5月世界卫生组织审议通过了《国际疾病分类第十一次修订本（ICD-11）》，ICD-11的传统医学章节中所使用的英语术语难以准确匹配到相应的传统医学术语，如太阴"Greater yin"国际上常译为"Tai yin"，很难与中医术语的含义匹配。同时，由于医学知识体系、历史文化的差异性，传统医学（以下简称"中医"）与现代医学（以下简称"西医"）在相同概念的匹配上存在知识特色宝藏，例如诺贝尔奖获得者屠呦呦发明的青蒿素受到东晋葛洪《肘后备急方》启发，证明了古代医学典籍中蕴含着丰富的专业知识可以"古为今用"。因此，通过术语映射手段建立中文术语与国际标准之间的对应关系，有助于将古代公共卫生知识进一步标准化、国际化，促进信息资源共享和知识服务。前人主要围绕中医学与西医学、口语化与国际疾病分类标准、中文与UMLS之间的术语映射展开研究。此外，在PubMed数据库中，生物医学信息检索者常用的一个主要检索类型是疾病名称（杨海锋，2015），疾病名称术语映射本质上体现了不同知识体系之间的融合与互补性。通过映射，可以将来自各体系的相关概念和术语整合在一起，形成一个更加全面和统一的疾病知识框架，从而提高知识的完备性和可共享能力。基于此，本文借助包含中西医疾病名称语句的语言规则构建同义词模式，进而挖掘同义词群。同时，将西医疾病名称作为隐性知识，采用语义计算方法实现中国传统医学与国际标准术语的关联映射，最终提高公共卫生领域术语的高质量建设和知识服务。


(c) 2024 中国中文信息学会计算语言学专业委员会

## 2 相关研究与理论基础

### 2.1 术语映射

#### 2.1.1 中西医疾病名称术语映射

现有研究主要是利用疾病名称的字面相似度、词向量训练、模式匹配等3种方法实现疾病名称映射。郭思成（2019）等以《中国中医药学主题词表》和《中文医学主题词表》为例，利用深度学习工具Word2Vec提出了一种在语义层面上进行自动化匹配计算的方法。姚涛等（2017）提出了一种基于映射字典学习的跨模态哈希检索算法，利用映射字典学习降低了算法复杂度。丁泽源等（2021）使用结合注意力机制的双向长短期记忆网络抽取中文生物医学实体间的关系。对于中西医病名之间字面相似度偏低的问题，如“核病”与“注气”、“霍乱”与“紧病”，传统的基于字面相似度的匹配算法难以起到较好的映射效果。尽管利用Word2Vec等机器学习技术训练得到的词向量可以通过计算余弦相似度来获取一定的匹配效果，但这种方法需要较长的训练时间，且对于低频词汇的效果欠佳。模式匹配法是从句法角度分析语料，依赖于自定义的抽取模式，从语料中进行语义关系抽取，属于传统的语义抽取方法之一。有学者提出了基于模式识别的中医药术语同义词提取方法，对中医药行业的基础学科名词术语进行规范化考订（Liao X，2020）。

模式匹配虽能依据事先定义的规则进行快速准确的信息提取，但其局限性在于人工定义规则无法覆盖所有可能的匹配模式。半监督学习是机器学习的基本方法之一，且在标记数据稀缺的情况下，利用未标记的数据可以显著提高模型的性能。王加楠（2017）提出了基于模式的远监督关系抽取算法,使用了基于核的机器学习算法来克服关键信息不突出、数据集线性不可分等问题。杜小坤(2015)提出了一种基于信息元的模式匹配方法,解决了结构化信息不够准确、缺少有效的描述形式、处理耗时等缺点。王丰（2019）提出了一种迭代优化的模式匹配方法IOSMA, 利用已经匹配成功的元素对优化模式匹配算法，使模式匹配算法随着迭代取得了更好的匹配效果。Snowball（Agichtein, Eugene and Luis Gravano，2020）是一种半监督的机器学习方法，提供了一种从文本文档生成模式和提取元组的方法。并通过计算抽取模式的置信度，评估Snowball算法所生成的模式和元组的质量，只有那些被认为“足够可靠”的元组和模式才会被算法保留，用于系统的后续迭代，有利于规避错误规则造成的语义漂移问题，因此，本文将模式匹配法与词向量相互结合，形成覆盖率和准确率较高的新方法。

#### 2.1.2 中英文术语映射

美国国立医学图书馆构建了生物医学领域的UMLS，对具有来源、格式、内容异构特点的医学词表统一术语格式，以解决相同概念在不同系统中名称不统一的问题，被广泛应用于医学健康领域的概念及概念关系识别、语义标注、语义消歧、本体构建等方面。刘娇等(2018)避开了外部词典和知识库,利用余弦相似度对中英韩三个语种的对齐语料库,建立了不同语种词汇间的对应关系。Chen L（2023）等通过系统地探索基于字符串、基于语义，以及字符串-语义组合的映射策略，来研究将中文医学实体映射到UMLS的技术边界。Dong Xin等（2022）提出了一种基于子网络的症状术语映射方法,有效地表示了临床症状术语，尤其是未记录术语的嵌入特征。Kim HK等（2020）结合字符串匹配方法和BioBERT 生成的术语嵌入向量，利用结构和上下文信息来计算源词和目标词之间的相似度量。Ruan T等人（2017）通过呼叫百度翻译，将中文医疗实体翻译成英文，计算平移和UMLS概念之间的Jaccard距离以进行映射。由于医学实体的多样性，翻译质量并不稳定，仅通过比较Jaccard的距离，很难其准确映射到UMLS上。随着深度学习的发展，预训练语言模型（PLMs），如BERT（Devlin J，2019）、GPT（2019）等，已被广泛应用于自然语言处理（NLP）的各个领域。预训练语言模型不仅可以学习医学术语的字符特征，同时可以获得与医学术语上下文相关的语义信息。刘方宇提出了一种对比度量学习方法来学习UMLS概念的自对齐，并训练了一种名为SapBERT的跨语言语言模型，为中英文疾病术语映射提供新的方案。由此可知，统一医学语言系统是一个相对完整的医学术语系统，对于医疗实体的规范化至关重要，但其主要由英语医学术语组成，对于英语以外的语言映射效果有待检验。

### 2.2 知识组织与知识发现

国内外对医学信息组织的相关研究主要围绕知识表示、知识聚合以及跨词表映射等。孙海霞（2013）等针对当前医学量表资源组织与服务粒度相对粗放问题，提出一种细粒度医学


(c) 2024 中国中文信息学会计算语言学专业委员会

量表文档知识表示框架，于诗睿（2024）等通过集成现有知识体系中的术语与新文献中的主题，使用上下文嵌入聚类主题模型对生物医学领域的专业术语进行层次化组织和映射。Saitwal H（2012）等将电子健康记录中的专有药物编码映射到SNOMEDCT和UMLS元词库中，建立一个多层次的分类系统，用于查询一个i2b2临床数据仓库。知识发现是从大量数据中获得有效、新颖、有潜在应用价值的和最终可理解的模式的高级处理过程（全国科学技术名词审定委员会，2019）。马捷（2024）等基于统计分析与关联规则的方法，挖掘中医用药规律与临床症状以及症—药规律。毛进（2024）等提出一种疾病知识网络表示学习模型及其链路预测算法,以进行疾病知识的关联关系挖掘与预测。Eissa等（2016）设计了粗糙集理论、人工神经网络、遗传算法和粗糙计量理论相结合的混合知识发现模型，通过在冠心病和丙型肝炎病毒数据集上的应用，展示了混合知识发现模型在提炼关键医疗指标和降低误诊率方面的潜力。

本文认为，对于中西医疾病名称映射，或是中英文疾病名称映射，均以数据驱动为核心，从各种辞典、文献资料中以机器学习的方式抽取同义词的过程。且在文本语料中，同义词与其连接词共同出现，可以将其归纳为根据关系规则获得疾病名称同义词对的过程。其目的在于向使用者屏蔽原始数据的繁琐细节，从原始数据中提炼出有效的、新颖的、潜在有用的同义词组。因此，本文以知识组织和知识发现为理论依据，利用半监督学习算法将分散分布的不同医学体系下的疾病名称建立联系，构建疾病名称同义词群，并利用预训练语言模型映射到国际疾病分类标准，实现古代专业知识的规范化和国际化，提高知识服务能力，在理论上是可行的。

## 3 模型构建

为了实现术语映射，本文设计了疾病名称映射框架图。主要包括3个方面，一是筛选目标数据，以中西医疾病术语映射、中英文术语映射为具体目标，采集相关度和专业程度较高的数据源，减少数据噪声；二是确定术语关联关系，挖掘关联关系和寻找内在规律的过程，针对医学疾病名称术语的字面语义关系较低，选择Snowball半监督学习算法挖掘关联关系，从而规避语义漂移问题；三是通过算法抽取的疾病名称同义词对构成同义词群，降低人工抽取的主观偏差。基于此，本文设计了如图1所示的疾病名称映射的主要框架。（1）采集多源数据：由于本

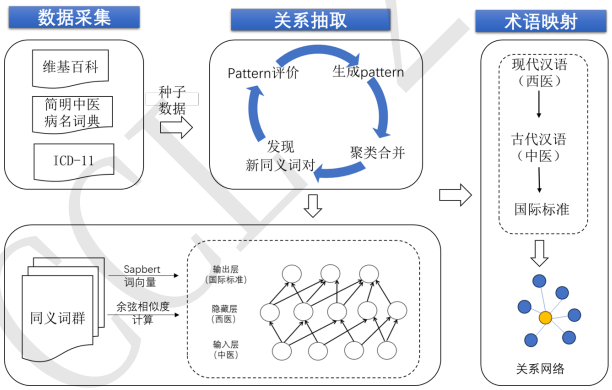



图 1. 疾病名称映射框架图

文涉及古代疾病名称、现代疾病名称、国际标准名称三者之间的相互映射，因此需要根据不同数据存储特点，完成数据采集任务。而《简明中医古病名辞典》参阅了历代有关中医古籍，辑录临床各科的古病名，并进行了中西医病名对照，可以从中抽取术语表达模式；ICD-11是世界卫生组织发布的最新版疾病分类系统，具有较高的国际地位和影响力，是本文国际标准术语映射目标；医学百科存在三种知识关联结构，即网状关联结构、层级关联结构和混合关联结构，包含疾病名称同义词、上位词、缩微词等信息，可以作为映射框架，将术语名称采用同义、近义、上下位等方式进行多种类型的映射。（2）中西医疾病名称映射：Snowball算法是一种从文本数据中抽取结构化信息的技术，其本质是一种半监督学习方法，即利用少量标注数据在未知语料中预测相关关系。《简明中医古病名辞典》内疾病名称之间的同义模式较为稳定，如句子“大瘕泄《难经·五十七难》大瘕泄者，里急后重，数至圊而不能便，茎中痛。即指由湿毒阻滞肠道所致腹痛、腹泻、里急后重、便下脓血者，相当于现代医学之细菌性痢疾”，且关系模式的重复率较高，基于此，本文利用Snowball算法完成疾病名称同义关系抽取，利用Snowball半


（c）2024 中国中文信息学会计算语言学专业委员会

监督学习算法，抽取中西医疾病名称同义关系模式，构建中西医同义词群。（3）中医向国际标准映射：西医疾病名称在国际上标准化程度较为完善，本文借助多层感知机的思想，将每一个中医疾病名称作为网络的输入层，每一个西医疾病名称作为网络的隐藏层，而国际标准疾病名称术语作为输出层，输入层不涉及任何计算，只需融合Sapbert生物医学预训练语言模型生成的Bert词向量，利用余弦相似度实现隐藏层和输出层的同义计算，即可完成中医向国际标准的映射。

### 3.1 疾病名称术语映射

#### 3.1.1 基于Snowball算法抽取疾病名称同义词对

利用算法Snowball算法抽取疾病名称同义词对主要流程如下：首先，将人工定义的一般性种子数据作为算法的输入部分，以此减少人工参与的程度，实现从文本中自动抽取结构化数据的目标。其次，利用该算法自定义的五元组（L，标签1，M，标签2，R），其中标签1和标签2代表疾病名称实体，L、M、R代表与疾病名称实体相关联的左、中、右的句子上下文。Snowball算法通过将句子上下文信息编码为带权重的向量来学习文本中的疾病名称同义关系，这些向量基于实体周围一定范围内的词语及其出现频率，以此衡量每个词语在特定上下文中的重要性，并据此抽取同义词对。例如，模式（[有时，0.2]，感冒，[又称，0.5]，伤寒），可以匹配“有时感冒又称伤寒”，其中“有时”和“又称”分别带有重要性权重0.2和0.5，从而帮助识别“感冒”和“伤寒”作为同义的疾病名称。该方法使得算法能够在仅有少量标注数据的情况下，从大量未标注的语料中有效地学习和扩展疾病名称的同义关系。然后，利用公式1得到不同五元组向量之间的相似度以及最小相似度的阈值。借助数学领域的内积概念，将五元组内“左”、“中”、“右”三个向量与其他元组向量进行乘法运算，以此衡量不同元组的相似度。其中匹配是指抽取的同义关系模式与文本中的具体实例相符合。例如，若模型抽取的同义关系模式为（[有时，0.2]，感冒，[又称，0.5]，伤寒），并且有一句文本是“鼠疫，又称，黑死病”，则认为该模式匹配这句话的内容。不匹配则指在关系模式匹配过程中，某个模式与文本中的具体实例不相符。例如，在某个关系模式下，如（[有时，0.2]，感冒，[又称，0.5]，伤寒），识别出感冒又称“鼠疫”，而在先验知识下，感冒又称“伤寒”而不是“鼠疫”，则（伤寒，鼠疫）同义词对被认为不匹配。

$$\operatorname{match}\left(t_p, t_s\right)=\begin{cases} l_p l_s+m_p m_s+r_p r_s, (\text{ 匹配}) \\ 0, \quad (\text{ 不匹配}) \end{cases} \tag{1}$$

由此可知，Snowball算法利用文本数据中的重复性和固有结构，通过迭代式的模式扩展过程，能够高效地从大量的文本数据中抽取出目标关系。此外，Snowball算法的优点在于能够自动调整抽取模式置信度，使挖掘的关联规则更加准确，即在每一次挖掘模式和同义词对的迭代过程中，仅让大于阈值的关系模式和实体对进入下一轮的迭代，从而保证最后生成的同义词对的质量。

#### 3.1.2 Bert词向量计算

Snowball算法在关系模型和同义词对抽取过程中，其核心是简单的规则匹配，而对于语义层面的同义词抽取效果欠佳，造成覆盖率不够、类推性不足等语义漂移问题。中医疾病名称向ICD-11映射的关键在于语法结构和表达习惯的差异性。虽然BERT模型在一定程度上能够处理未知词汇，但对于某些古代疾病名称的相关词汇难以生成准确的词向量。本文利用Snowball算法抽取的古今疾病名称同义词对，以现代医学疾病名称为中介桥梁，利用Bert词向量表征语义关系，从而从语义角度提高疾病名称映射准确性，克服单纯依靠规则匹配造成的语义漂移问题。

SapBERT是一种用于生物医学实体链接任务的开源预训练模型。与传统的BERT模型相比，SapBERT在预训练阶段通过自我对齐目标来学习生物医学实体的表示，将模型的同一实体在不同语言或领域中的不同描述或表示对齐到一个共同的表示空间，有利于处理跨语言或跨领域的任务，从而更好地理解实体之间的关系。此外，利用余弦相似度公式量化向量间的距离，如公式2所示，该值的范围为[-1,1]，-1为完全不相似，1为完全相似。因此，本文基于嵌入表示

来衡量中文疾病名称与国际标准之间的语义相似性，以此促进医学术语的规范化和互操作性。

$$similarity(A,B)=\frac{AB}{||A||\times||B||}=\frac{\sum_{n}^{i=1}A_i\times B_i}{\sqrt{\sum_{n}^{i=1}A_i^2}\times\sqrt{\sum_{n}^{i=1}B_i^2}} \tag{2}$$

## 3.2 术语映射评价指标

### 3.2.1 置信度

通过评估同义关系模式以及元组的置信度筛选同义词对，利用公式2计算在模式P下识别正确同义词对与错误同义词对的个数统计。若处于“xx又称xx”模式下，可以在4条语句中识别正确，1条语句中识别错误，则该模式的置信度为Conf(P)=4/5=0.8。在对关系模式进行置信度评价后，会得到若干个候选同义词对。对于每个同义词对，存在若干模式集合以及该元组句语中上下文和匹配函数的相似性度量，如图2所示。

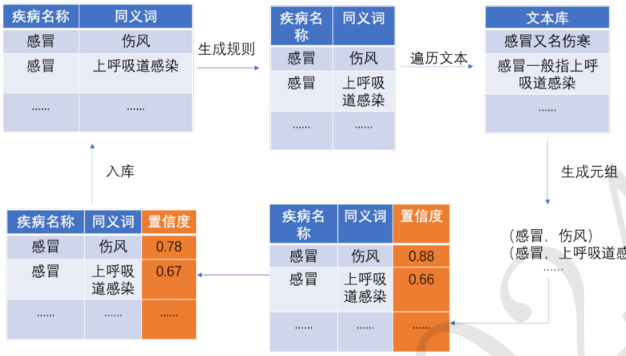



图 2. 疾病名称同义词关系抽取流程

因此，利用公式3和公式4按对应模式和上下文的匹配程度来表示计算元组置信度，且模式和上下文之间的匹配程度越低，产生无效同义词对的机会就越高。若同义词对周围的上下文“足够相似”，即同义词对相似度达到阈值，则根据语料库中出现的已知元组进行聚类，进而形成关系抽取模式。最后，将形成的关系抽取模式运用到语料库里面提取更多的同义词对，不断重复上述过程。同时将置信度较高的预测结果加入到种子数据中，作为新的训练数据，重新训练模型。

$$Conf\,(P)=\frac{P(right)}{P(right)+P(error)} \tag{3}$$

$$Con(T)=1-\prod_{i-0}^{|p|}(1-(conf(pi)\times Match(ci,pi))) \tag{4}$$

### 3.2.2 hit@k指标测度

由于国际疾病分类（ICD-11）标准是根据病因、病状等因素进行综合分类，而语言风格、文化背景以及医学知识的差异性，中英文疾病名称一对一、一对多或者零对应的映射关系，因此常见的测评指标如精确率、召回率、F1值等很难客观评价映射准确性，更适合采用相关性评估进行评测。hit@K，也称为Top-K准确率，是一个衡量推荐系统性能的指标，它表示在所有搜索查询中，系统在前K个推荐结果中至少正确推荐了一个相关项目的比例，该指标常用于评估推荐系统的准确性。如公式5是hit@K指标的计算方法，其中分子是所有满足条件的查询数量（即前K个推荐中至少有一个是正确推荐的），分母则是所有搜索查询的总数，K值可选取1、3、10等。

$$hit@K=\frac{\sum_{q\in Q}isCorrect(q,K)}{|Q|}\% \tag{5}$$

因此本文通过近似匹配的原则，利用余弦相似度的算法得到中文疾病名称与多个英文疾病名称的若干个余弦相似度，并引入hit@k指标计算其有效性，即在Sapbert模型给出的前k个余弦相似值中，正确映射占所有映射结果的百分比。如果Sapbert预训练语言模型计算余弦相似度的正确映射排名越靠前，说明该方法完成中英文映射效果较好。

CCL 2024

# 4 实验

## 4.1 数据采集与预处理

本文选取传染性疾病作为实验语料，首先，从维基百科获取传染性疾病词条信息。然后，将以图片形式存储的《简明中医古病名辞典》利用OCR识别工具转换为文字，在此基础上采集包含“传染”的相关疾病描述，共计125个疾病名称。如图3所示，将上述采集的语料按照HTML标签，以人工方式进行命名实体识别（使用NAM表示疾病名称、BOOK表示疾病名称出处），经过去重整理后得到条165条文本数据。最后，在世界卫生组织官方网站下载《国际疾病分类》第11次修订本（简称ICD-11）作为为跨语言映射的基准，剔除非疾病术语，如药品名称、体征和症状等，得到2478个国际标准术语。

<NAM>小腿慢性湿疹</NAM>。治宜疏风解毒祛湿，方用草薢渗湿汤加味。
<NAM>下痞</NAM> 《华佗神医秘传》卷16: “华佗治下府神方。” 即<NAM>秽疮</NAM>。详见该条。下症疮《景岳全书》卷64: “治痔漏下痈疮。” 即秽疮。详见该条。
<NAM>下格</NAM> 《医钞类编》卷15: “……此为下格。” 即指肠内因秽浊，恶物结滞，阻碍大便传导所致大便数旬不通，时时作呕，饮食不尽，食入即吐等病证。相当于现代医学的便秘、肠梗阻等症。治宜荡涤肠垢，方用小承气汤。
<NAM>下注疮</NAM> 《世医得效方》卷19: “名曰下疳疮。” 即<NAM>烂腿疮</NAM>。详见该条。
<NAM>下部病</NAM> 《心印绀珠经》下卷: “瘙疝之病俗云下部病。” 即<NAM>狐疝</NAM>。详见该条。下部愚 《别录》: “苦参，……(治)恶疮、下部慝。” 即阴疮。详见该条。
<NAM>下消</NAM> 《丹溪心法》: “……下消。” 又称肾消、消肾。指因肾水亏竭，蒸化失常

图 3. 部分语料示例

## 4.2 传染性疾病中西医疾病名称映射

依托于GitHub内开源Snowball半监督学习模型，将源代码下载到本地，利用PyCharm编译器完成如下实验步骤:

（1）初始化snowball算法，根据“二八定律”经验值，在许多情况下大约80%的结果通常来自于20%的原因。因此，本文随机抽取22个可以反映出待挖掘关系的样本种子，如同义关系（丁疮；疔疮）、上下位关系（食积；九积）作为snowball算法的种子数据，如表1所示。设置迭代轮次10次，规则置信度为0.6，公式1同义词对匹配函数阈值为0.6。

| 疾病名称 | 别称 | 疾病名称 | 别称 |
|---|---|---|---|
| 丁疮 | 疔疮 | 儿晕 | 儿痉 |
| 七情饥饱款 | 七情伤感嗽 | 九子疱 | 颈淋巴结结核 |
| 八脚虫疮 | 阴虱疮 | 九道出血 | 大衄 |
| 儿风 | 妊娠风痉 | 刀癣 | 泛发性神经性皮炎 |

表 1. snowball算法部分种子集

（2）利用自举（bootstrapp）方法初始化迭代。在每次迭代中，根据步骤（1）设置的初始化阈值，Snowball算法尝试在文本语料中寻找与种子数据相似的模式，然后将这些模式和它们对应的关系实例加入到种子集合中。随着迭代的进行，种子集合会逐渐增大，算法能够识别的关系模式也会越来越丰富。

（3）在自举迭代过程结束后，Snowball算法会将所有提取的关系写入到输出文件中，该文件包含了每个关系实例的详细信息，如实体对、它们之间的关系、以及关系的置信度等。本文在迭代轮次为10次，规则置信度及同义词对阈值为0.6的条件下，得到21条规则，如表2所示。

（4）以现代疾病名称为入口，整理与其具有较高相关性的古代疾病名称形成同义词群，如狂犬病为现代疾病名称，与之相对应的古代疾病名称有颠犬伤、狂犬伤、狂犬啮人、疯犬咬伤、犬伤、狗啮疮、风犬伤毒、风狗咬、狗啮。经噪声整理后，去除错误规则，最终提取出12个古今同义词关系模式，如表4所示，并将算法抽取的规则划分为等同关系、等级关系以及相关关系。其中等同关系频次最高，相关关系频次最低。共计64条古代疾病名称可以映射到现

| 序号 | 同义词对 | 置信度 | 规则 |
|---|---|---|---|
| 1 | （八脚虫疮，阴虱疮） | 0.81 | (……即……) |
| 2 | (干血痨，干血) | 0.79 | (……称……) |
| 3 | (白缠喉，白喉) | 0.67 | (……相当于现代医学的……) |
| 4 | （禽流感病毒，甲型流感病毒） | 0.63 | (……属……) |
| 5 | （百日咳，鸡咳） | 0.66 | (……俗称……) |

表 2. Snowball算法生成的部分规则

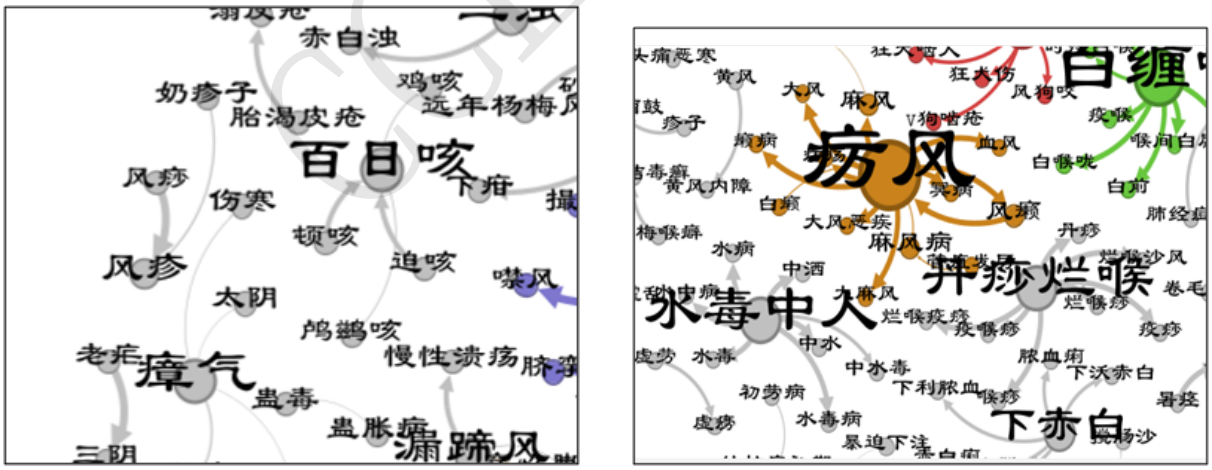
代疾病中，而34条古代名称未能与现代疾病名称相对应，且存在多个古代疾病名称表示一个现代名称的情况，如现代结核病大致相当于于丁吴痞、尸注、初劳病、抱儿痨、注气、虚劳。而伏气、后不利、血风等未找到与现代疾病名称的映射结果。

| 关系 | 规则 | 频次 |
|---|---|---|
| 等同关系 | 又称、又名、亦指、是一种 | 64 |
| | 即、简称、名为、指、俗称、俗名 | 43 |
| 等级关系 | 属于、之一、属 | 15 |
| 相关关系 | 类似、如、参见 | 9 |
| | 病因、故、引起、多由、故又称 | 6 |

表 3. 疾病名称按不同关系统计频次

### 4.2.1 构建中西医疾病名称同义词群

部分疾病名称未能识别或识别错误，情况主要分为以下2类：一是疾病名称以符号连接，表示一般性停顿，以致未能识别其他同义词对。如语句“赤白痢，又称脓血痢，小肠泄，下利脓血，下沃赤白”中，“赤白痢”和“小肠泄”以及“赤白痢”和“下利脓血”之间没有连接词，导致未识别，说明对多个统一术语的抽取能力还有一定局限性。二是语义误差，即存在同义关系模式，但其语义表达不符合抽取要求。如语句“由于风疹的疹子来得快，去得也快，如一阵风似的，“风疹”也因此得名。”描述了疾病命名来源，但错误识别为“疹子…如…一阵风”。




(a) 疾病名称离散分布　　(b) 疾病名称集中分布

图 4. Gephi古代疾病名称可视化效果

为了进一步揭示疾病名称的关联性，本文利用Gephi进行可视化展示中西医疾病名称同义词群，其中节点代表疾病名称，边代表疾病名称的关系之间的关系，并按照其出现的频次作为该关系的权重，以表示关系的重要性。平均加权度是指整个网络中节点的度数以及节点之

间边的权重，反映了各个节点之间直接关联的程度。在该网络中，平均加权度为33，表明疾病名称之间存在强而紧密的联系。利用Gephi内置的Modularity聚类算法展示疾病名称的同义词群，并按颜色划分不同词群，由图4可知疠风、白缠喉、时毒等聚类效果较为明显，表明利用Snowball算法抽取疾病名称同义词对效果显著。此外，由网络分布的离散状态，推测疾病名称相关性较高，且疾病名称分布越集中，表明其涵盖的知识越粗糙，反之细粒度较高。如图4所示。

### 4.2.2 中西医疾病名称历时演变

由于《简明中医古病名辞典》内涵盖疾病名称的来源信息，如“丁奚《诸病源候论》卷47，小儿丁奚病者”。基于此，本文以“霍乱”、“结核病”、“猩红热”、“疟疾”和“风疹”5个中医疾病名称为例，根据《简明中医古病名辞典》内这5个疾病名称的出处典籍以及4.2.2节构建的中西医疾病名称同义词群，以疾病名称所在典籍成书时间作为节点，反映疾病名称在不同时代的演变过程和知识脉络。并利用镝数图表在线平台，构建的疾病名称历时变化桑基图。其中第一列为西医疾病名称，第二列为中医疾病名称，第三列为中医疾病名称选自的医学典籍，第四列为医学典籍的成书年代。如图5所示。

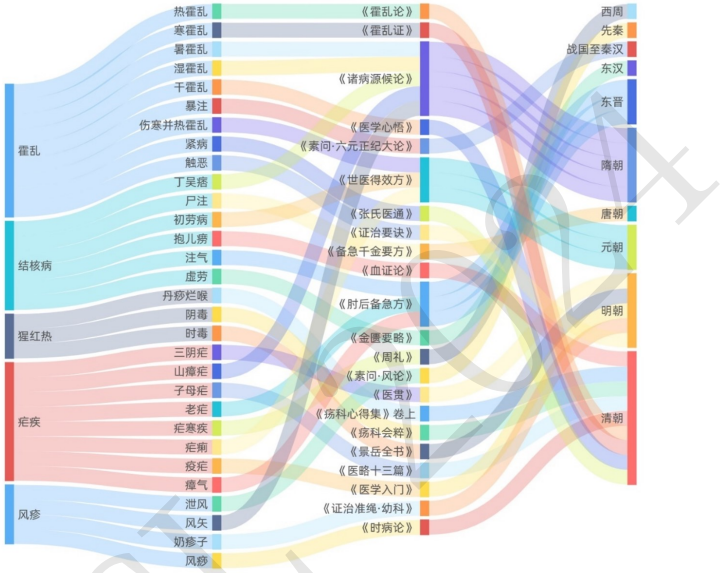



图 5. 疾病名称历时演变

图5从关联性、成熟度和延展性3个角度分析中西医疾病名称历时演变，对知识的关联性和时序进行更直观的揭示和溯源。一是从中西医疾病名称的关联性可知，两种医学体系命名相互融合。如西医称为“霍乱”，从战国至秦汉中医体系下将其依次命名为“暴注、湿霍乱、暑霍乱、触恶、热霍乱、干霍乱”，用户通过模糊检索“霍乱”即可得到大部分与之相关联的疾病名称。二是从医学典籍成书时间终止点明晰疾病发展的成熟度。以西医“结核病”为例，由图4可知自东汉至清代均有对该种疾病的医学记载，与其关联的中医疾病名称先后出现了“注气—丁吴痞—初劳病”的名称变化。本文结合辞典注释进行分析，由于古代对该种疾病症候较为模糊，使用注气这一泛化的名称表达对气滞症状的初步理解。随着对病理机制的不断加深，以丁吴痞命名，不断细化此类疾病，其中丁表示疼痛，吴表示胀满，痞表示阻塞。由此可知，在明朝时期结核病已形成较为完整的治疗体系。三是以节点密集程度衡量疾病名称的延展性。从典籍节点密集程度来看，《诸病源候论》中涉及霍乱、疟疾、风疹三种疾病名称表明该部典籍蕴含综合性知识。因此，典籍节点越密集知识覆盖面越广，典籍节点越稀疏表明其对某种疾病的专业度越强。从时间节点密集程度来看，隋朝至元代时期古代医学较为发达或处在疾病爆发时期。且从西周至清朝时期，节点大小呈递增趋势，表明医学知识逐渐丰富，医学典籍数量逐渐增多。

通过对疾病名称的溯源，一方面便于医学典籍的知识表示和智能应用，为用户提供更多的入口词，另一方面可以反映古代疾病名称的时间演化，进而归纳疾病的历史演变历程，技术的成熟度以及疾病的延伸情况，为用户推送更具时序逻辑关联性的知识内容与服务。


(c) 2024 中国中文信息学会计算语言学专业委员会

### 4.3 中医传染性疾病向国际标准映射

由4.1和4.2节，本文构建了中西医传染病疾病名称同义词群。基于此，将西医疾病名称和ICD-11国际标准术语转化为词向量的形式，采用余弦相似度计算的方法实现中国传统医学与国际标准术语的关联映射，主要流程如下：首先，加载SapBert预训练语言模型，使用tokenizer对ICD-11疾病分类术语集进行编码，并将其转换为模型的嵌入向量，从而得到2478个疾病术语的向量表示，将其以数组形式存储在numpy文件中。然后，由4.2.2节构建的中西医疾病名称同义词群，本文将西医疾病名称输入到Sapbert预训练语言模型中，利用python编程语言定义query的变量，输入待映射中文疾病名称，且同上所述将其转换为向量表示，作为查询向量。最后，调用利用ICD-11生成的numpy文件，利用公式2计算查询向量与所有ICD术语嵌入向量之间的余弦相似度，并按余弦相似度值降序排列。若余弦相似度排在第一位的结果即为正确映射，则hit@1为100%。而真实情况是在余弦相似度排在第5个甚至第10个预测时才会出现正确映射结果，因此本文依次评估hit@5和hit@15，表示Sapbert模型在前5个和前10个余弦相似度中出现真实值的输出占所有疾病名称的百分比。

| 序号 | 现代疾病名称 | ICD-11标准 | 模型预测结果 | 余弦相似度 |
|---|---|---|---|---|
| 1 | 鼠疫 | plague | plague | 0.924 |
| | | | bronchitis | 0.925 |
| | | | kleptomania | 0.926 |
| | | | trance disorder | 0.926 |
| | | | rabies | 0.927 |
| | | | dyspnea disorder | 0.928 |
| 2 | 霍乱 | Cholera | pulmonary oedema | 0.932 |
| | | | syngamosis | 0.932 |
| | | | pulmonary eosinophilia | 0.933 |
| | | | asthma | 0.933 |
| | | | pica | 0.935 |
| | | | adjustment disorder | 0.936 |

表 4. 部分疾病名称中英文映射

如表4所示，展示了部分中英文疾病名称Sapbert映射的余弦相似度值。以人工核验的方式，按照中文版ICD-11国际标准进行校对。利用公式5计算中英文疾病名称映射实验结果，达到23%的hit@1，42%的hit@5以及61%的hit@10。实验结果表明，在利用Sapbert这一预训练语言模型的强大基础上，并将国际疾病分类第十一版（ICD-11）的标准化术语整合进Sapbert模型中，能够提高中英文疾病术语之间的精准映射。

## 5 研究总结

本文针对疾病名称古今不对应、国际化标准化不对应这两个关键问题，提出了融合半监督学习与语义计算的疾病名称跨语言映射方法，能够将模式规则和词向量结合，自动挖掘古今疾病名称对应关系，有助于形成可信、可解释的知识表示与知识发现，推动构建人机两用的专业知识库，提高知识发现效率。最终实验结果表明，hit@10映射结果达到0.61，证明该方法切实可行。其优点在于，ICD-11结构简单、规范的特点，保证了语义映射机制和映射过程的规范性，提高了多知识组织系统集成的映射效率。通过映射能够快速提取“传统医学—现代医学—国际标准”的专业知识关联性，分别从语义关联性和时间演变两个维度建立知识的关联网络，有助于提高信息质量和服务效率。本文采用专业词典和维基百科等数据进行“富知识数据”进行抽取，获取疾病的古今疾病名称对应关系，准确性高但覆盖率还有待检验。今后可以重点对大规模古籍文本语料、古代医案等数据进行知识挖掘和建立关联关系，以更好的反映知识全貌。

## 参考文献

王加楠,鲁强. 2017. 基于模式的远监督关系抽取算法, 中文信息学报,31(02):122–131.


(c) 2024 中国中文信息学会计算语言学专业委员会

丁泽源,杨志豪,罗凌. 2021. 基于深度学习的中文生物医学实体关系抽取系统, 中文信息学报,35(05):70-76.

刘娇,崔荣一,赵亚慧. 2018. 基于共现词映射的中英韩跨语种文档相似度计算, 中文信息学报,32(03):55-63.

赵洁,贾君枝,司莉. 2023. 基于社会化问答的用户健康信息需求概念体系构建——以糖尿病为例, 情报理论与实践,46(07):150-157.

杨海锋. 2019. 用户行为在信息检索中的研究现状及发展动态评述, 图书情报知识,(06):79-88.

郭思成,李纲,周华阳. 2019. 基于*Word2Vec*的医学知识组织系统互操作研究——以词表间语义映射为例, 情报理论与实践,42(09):160-165+176.

杨海锋. 2019. 用户行为在信息检索中的研究现状及发展动态评述, 图书情报知识,(06):79-88.

姚涛,孔祥维,付海燕,等. 2018. 基于映射字典学习的跨模态哈希检索, 自动化学报,44(08):1475-1485.

王丰,王亚沙,赵俊峰,等. 2019. 一种基于迭代的关系模型到本体模型的模式匹配方法, 软件学报,30(05):1510-1521.

Kim HK,Choi SW, Bae YS. 2020. *A Context-Aware Term Mapping with String Matching and Embedding Vectors*, *APPLIED SCIENCES-BASEL*,10(21).

Dong Xin,Zheng Yi,Shu Zixin. 2020. *TCMPR: TCM Prescription Recommendation Based on Subnetwork Term Mapping and Deep Learning.*, *APPLIED SCIENCES-BASEL*,10(21).

Liao X,Bu Y,Jia Q. 2020. *Traditional Chinese medicine as supportive care for the management of liver cancer:past,present,and future*, *Genes&Diseases*,7(3):370-379.

杜小坤,李国徽,王江晴,等. 2015. 基于信息元的模式匹配方法,软件学,26(10):2596-2613.

Agichtein, Eugene and Luis Gravano. 2000. *Traditional Snowball: extracting relations from large plain-text collections.*, *Digital library.*

Chen L, Qi Y, Wu A. 2023. *Mapping Chinese Medical Entities to the Unified Medical Language System.* , *Health Data Science.*

Ruan T,Wang M,Sun J,Wang T, Zeng L, Yin Y, Gao J. 2017. *Interoperability and mapping between knowledge organization systems: Meta thesaurus—unified medical language system of the National Library of Medicine* , *Journal of Biomedical Semantics.*

Devlin J,Chang M-W,Lee K, Toutanova K. 2019. *BERT: Pretraining of deep bidirectional transformers for Language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human language technologies*, *Journal of Biomedical Semantics.*

孙海霞,郝洁,郭臻. 2023. 基于知识元的细粒度医学量表文档知识表示框架构建, 数字图书馆论坛*l informatics*, 19(12):86-98.

于诗睿,李爱花,杨雪梅. 2024. 融合知识组织体系的层次化主题挖掘方法研究, 数据分析与知识发现, 1-18.

Saitwal H,Qing D, Jones S, et al. 2012. *Cross-terminology mapping challenges: a demonstration using medication terminological systems*, *Journal of biomedical informatics*, 45(4),613-625.

毛进,侯博文,王依蒙. 2024. 基于知识元引用网络的细分领域演化特征研究, 情报理论与实践,47(02):107-115.

全国科学技术名词审定委员会. 图书馆·情报与文献学名词, *https://www.termonline.cn.*

Mohammed M. Eissa, Mohammed Elmogy, Mohammed Hashem. 2016. *Rough-Granular Computing knowledge discovery models for medical classification*, *Egyptian Informatics Journal*,17(3):265-272.

CCL 2024

# 中文语法纠错的多轮解码方法研究

**王晓盈, 穆玲玲*, 许鸿飞**
郑州大学计算机与人工智能学院，河南省 郑州市，450001
xywangnlp@qq.com; iellmu@zzu.edu.cn; hfxunlp@foxmail.com

## 摘要

在语法纠错（Grammatical Error Correction, GEC）任务上，序列到序列（Sequence-to-sequence, seq2seq）模型与序列到编辑（Sequence-to-edit, seq2edit）模型相比可以取得相当或更好的性能。序列到编辑模型通常通过多次迭代解码，而序列到序列模型则以从左到右的方式一次性解码，不考虑后续的词语。通过在序列到序列模型中应用多轮解码（Multi-Turn Decoding，MTD）来迭代改进前一轮的修正结果，可能会进一步提升性能。然而，多轮解码会增加推理的计算成本，且前一轮修正中的删除或替换操作可能会导致原始输入中有用的源语句信息丢失。本文提出了一种早停机制来提高效率。同时，为解决源语句信息丢失问题，本文将原始输入与上一轮的修正结果合并为一个序列。在NLPCC2018测试集、FCGEC验证集和NaCGEC测试集的实验结果表明，本文方法可在BART基线上能带来一致且显著的性能提升，F0.5值分别提高了+2.06，+2.31和+3.45，分别取得了47.34，54.58和62.09的F0.5值。

**关键词：** 语法纠错 ； 多轮解码 ； 序列到序列

# Multi-Turn Decoding for Chinese Grammatical Error Correction

**Xiaoying Wang,Lingling Mu*, Hongfei Xu**
School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou Henan, 450001
xywangnlp@qq.com; iellmu@zzu.edu.cn; hfxunlp@foxmail.com

## Abstract

Sequence-to-sequence (seq2seq) models achieve comparable or better Grammatical Error Correction (GEC) performance compared to sequence-to-edit (seq2edit) models. Seq2edit models normally iteratively refine the correction result, while seq2seq models decode only once in a left-to-right manner unaware of subsequent tokens. Iteratively refining the correction results from the previous round with seq2seq models via Multi-Turn Decoding (MTD) may lead to better performance. However, MTD increases the inference costs. Deleting or replacing corrections in previous rounds may lose useful information in the source input. This paper presents an early-stop mechanism to alleviate the efficiency issue. To address the source information loss issue, this paper proposes to merge the source input with the previous round correction result into one

*通讯作者
©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

sequence. Experiment results on the NLPCC2018 test set, FCGEC development set and NaCGEC test set show that our approach can lead to consistent and significant improvements over the strong BART baselines (+2.06, +2.31 and +3.45 F0.5 scores respectively), and obtaining F0.5 scores of 47.34, 54.58 and 62.09.

**Keywords:** Multi-Turn Decoding , Grammatical Error Correction , Sequence-to-sequence

## 1 引言

语法纠错（Grammar Error Correction，GEC）旨在纠正给定句子中的语法错误[Ng et al., 2013, 2014]。目前，主流的语法纠错方法分为序列到编辑和序列到序列两种。序列到编辑方法将语法纠错任务视为序列标注任务，通过预测每个词语的编辑标签（如保留、删除、插入、替换等）并迭代处理，直至所有词语均被标记为保留[Malmi et al., 2019, Stahlberg and Kumar, 2020, Omelianchuk et al., 2020, Yuan et al., 2021]。这种方法往往需要多次校验以达到准确修正。相对地，序列到序列方法将语法纠错任务视为机器翻译任务，即从有语法错误的句子转换到语法正确的句子[Zhao et al., 2019, Kiyono et al., 2019, Wang et al., 2021, Li et al., 2022a, Fang et al., 2023]。序列到序列模型对输入句子进行编码，并自回归解码出修正后的句子来实现这一转换[Omelianchuk et al., 2020, Liu et al., 2021a]。为提升性能，现有方法通常采用预训练模型来提高纠错性能，例如序列到编辑方法使用BERT [Devlin et al., 2019]和XLNet [Yang et al., 2019]，序列到序列方法使用BART [Lewis et al., 2020]和T5 [Raffel et al., 2020]。

序列到序列模型在未使用特定语言编辑操作的情况下，表现出与序列到编辑方法相匹配或更优的性能。然而，当前的序列到序列语法纠错研究通常仅执行一次解码，忽略了后续词语的影响。多轮解码（Multi-Turn Decoding，MTD）策略可以通过迭代优化提升性能。训练多轮解码模型以生成给定错误解码结果的参考结果，可能通过自我修正机制增强学习效果。

多轮解码不直接采用纠错模型单轮的解码结果作为最终的纠错结果，而是通过将上一轮的修正结果作为新一轮的输入，要求模型产生新的纠错结果，并迭代修正，直至连续两轮的解码结果相同。尽管这种方法提升了性能，但增加了推理的计算成本。为解决效率问题，本文提出一个与序列到序列模型协同工作的早停机制。该机制以句子结束标记（<eos>）的隐藏表示作为输入，并在以下两种情况下终止多轮解码：1）下一轮的修正结果与当前修正结果一致，或2）与当前修正结果相比，下一轮的修正结果与参考结果的编辑距离更大。

此外，在多轮解码过程中，若模型在前几轮中删除或替换了某些词语，那么后续的纠正将遗失这些词语的信息，即使它们可能为后续纠正提供有价值的信息。为解决信息丢失问题，本文提出同时将源语句与上一轮的修正结果同时输入纠错模型，直接将两个序列拼接[Correia and Martins, 2019]输入预训练模型会使长度翻倍，且不利于模型对比分析两个序列中对应位置的词语。双编码器结构[Denkowski, 2015]会增加模型的复杂度，不利于对预训练模型语言能力的充分利用。本文提出基于序列对比的源信息融合方法，利用两个序列相近的特性，对比两个序列中的相同片段和差异片段后再依序拼接，避免相同片段的重复输入。然后利用编辑标记或位置编号对混合序列中每个词语在两个序列中的角色予以区分。

本文的主要贡献如下：

- 本文针对现有的序列到序列纠错方法仅进行一次从左到右的解码，在解码过程中不能考虑后续词语的问题，提出通过多轮解码来重复检查纠正现有的纠错结果，改进序列到序列语法纠错方法。
- 为提升多轮解码的效率，本文提出早停机制，在下一轮可能不会产生更好的纠错结果时，提前结束多轮解码。
- 本文提出源信息融合方法来解决多轮解码中删除和替换词语导致的信息丢失问题，并根据两个序列相近的特性，提出基于对比的序列混合方法，避免序列拼接导致的效率问题。
- 在NLPCC2018测试集、FCGEC验证集和NaCGEC测试集上，本文的方法相对于强大的BART基线带来了+2.06，+2.31和+3.45的F0.5值提升，分别取得了47.34，54.58和62.09的F0.5值，证明了本文方法的有效性。


(c) 2024 中国中文信息学会计算语言学专业委员会

**算法 1** 多轮解码与早停机制。输入：输入纠正句子$x$，中文语法纠错模型$M$，提前停止分类器$C_e$，最大解码轮数$n$，提前停止阈值$\tau$；输出：更正后的句子$y$。

$\hat{x}_0, h_{d,0}^{<eos>} = M(x)$;
$p_e = C_e(h_{d,0}^{<eos>})$;
**if** $p_e > \tau$ **then**
  $y = \hat{x}_0$
**else**
  **for** $t = 1$ to $n$ **do**
    $\hat{x}_t, h_{d,t}^{<eos>} = M(x, \hat{x}_{t-1})$;
    $p_e = C_e(h_{d,t}^{<eos>})$;
    $y = \hat{x}_t$
    **if** $\hat{x}_{t-1} == \hat{x}_t$ or $p_e > \tau$ **then**
      break;
    **end if**
  **end for**
**end if**
**return** $y$

## 2 基于序列到序列的中文语法纠错

序列到序列模型$M$接收待纠正的输入序列$x$，并输出修正后的序列$\hat{x}$。该模型由编码器和解码器两部分构成。

编码器接收输入序列$x$，并计算上下文隐状态向量$h_e$：

$$h_e = encoder(x) \tag{1}$$

解码器根据编码器的隐状态$h_e$和已生成的解码序列$\hat{x}^{<k}$来生成当前的隐状态$h_d^k$：

$$h_d^k = decoder(h_e, \hat{x}^{<k}) \tag{2}$$

在解码过程中，$\hat{x}^k$代表序列中的第$k$个标记，$\hat{x}^0$对应于序列的开始标记<sos>，而$\hat{x}^{<k}$则表示从$\hat{x}^0$到$\hat{x}^{k-1}$的标记序列。解码器中的分类器利用隐状态$h_d^k$来预测词汇表中每个词语的概率，并选择概率最高的词语作为下一个解码步骤的$\hat{x}^k$。

$$\hat{x}^k = classifier(h_d^k) \tag{3}$$

解码器将持续执行此过程，直至分类器在接收到隐状态$h_d^{<eos>}$时输出序列结束标记（<eos>）。

通过重构损坏文本进行预训练，大规模语料库的知识可被编码进模型参数。损坏文本是指从原始文本中选择部分词进行mask或换序等操作后的句子。在语法纠错任务上对预训练模型进行微调（如BART和T5）可进一步提升性能[Sun et al., 2021, Rothe et al., 2021]。

## 3 本文方法

### 3.1 结合早停机制的多轮解码

在语法纠错任务中，序列到序列语法纠错模型$M$输入可能包含语法错误的句子$x$，并输出修正后的句子$\hat{x}$。本文不直接采用$\hat{x}$作为最终输出，而是通过将上一轮的修正结果$\hat{x}_{t-1}$作为新一轮的输入，要求模型将其修正为$\hat{x}_t$，以此迭代地进行修正，直至$\hat{x}_t = \hat{x}_{t-1}$。迭代的终止条件是连续两次解码产生相同的序列。尽管这种方法提升了性能，但也增加了推理的计算成本。为解决效率问题，本文训练了一个与序列到序列模型协同工作的早停机制。

早停机制包括一个轻量级的逻辑回归分类器$C_e$，用于预测提前停止多轮解码的概率$p_e$。$C_e$由权重向量$w_e$和偏置$b_e$组成。在解码$\hat{x}_{t-1}$时，本文使用特殊句子结束标记（<eos>）的隐状态表示$h_{d,t-1}^{<eos>}$来计算早停的概率：


(c) 2024 中国中文信息学会计算语言学专业委员会

| | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 原始文本（$x$）： | 我 | 们 | 明 | 天 | 从 | 石 | 家 | 庄 | 出 | 发 | 去 | 北 | 京 | | | 中 | | 科 | | 院 | 。 |
| 前一轮解码结果（$\hat{x}_{t-1}$）： | 我 | 们 | 明 | 天 | 从 | 石 | 家 | 庄 | 出 | 发 | 去 | | | 河 | 北 | 中 | 国 | 科 | 学 | 院 | 。 |
| 混合序列： | 我 | 们 | 明 | 天 | 从 | 石 | 家 | 庄 | 出 | 发 | 去 | 北 | 京 | 河 | 北 | 中 | 国 | 科 | 学 | 院 | 。 |
| 编辑标记： | e | e | e | e | e | e | e | e | e | e | e | d | d | i | i | e | i | e | i | e | e |
| 位置编码： | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 0 | 0 | 14 | 0 | 15 | 0 | 16 | 17 |
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 0 | 0 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 |

图 1：基于序列对比的源信息融合。对于编辑标记，“e”、“d”和“i”分别表示相等、删除和插入。“0”表示位置编码中的填充。

$$p_e = \sigma(h_{d,t-1}^{<eos>} \bullet w_e + b_e) \tag{4}$$

其中“$\bullet$”和“$\sigma$”分别表示向量点积和sigmoid激活函数。

本文使用$p_e$和早停标签$y_e$之间的二元交叉熵（Binary Cross Entropy, BCE）损失来训练早停分类器：

$$l_e = \mathrm{BCE}(p_e, y_e) \tag{5}$$

在多轮解码训练中，本文首先解码得到本轮的解码结果$\hat{x}_t$，随后根据前一轮的解码结果$\hat{x}_{t-1}$、本轮的解码结果$\hat{x}_t$和参考答案$r$来标记早停分类器的分类标签$y_e$，如果：1）本轮的解码结果$\hat{x}_t$与前一轮的解码结果$\hat{x}_{t-1}$相同，或2）参考答案$r$与本轮解码结果$\hat{x}_t$之间的编辑距离大于与前一轮解码结果$\hat{x}_{t-1}$之间的编辑距离，则分类标签$y_e$为真。引入第二个编辑距离条件旨在确保多轮解码不会降低性能。

训练损失是原始序列到序列生成损失$l_{seq2seq}$和编辑损失$l_e$的加权组合：

$$l = l_{seq2seq} + \lambda * l_e \tag{6}$$

本文在推理阶段采用算法 1。设定最大解码轮数$n$为3，若$\hat{x}_t = \hat{x}_{t-1}$或$p_e > \tau$，则触发早停。参数$\lambda$和$\tau$分别默认设为1和0.5。将$\lambda$设为1意味着在训练过程中，纠正任务和早停分类器被赋予同等重要性。阈值$\tau = 0.5$表示当停止解码的概率超过0.5时提前终止多轮解码。对于二分类任务，$\tau = 0.5$是一个合理的选择。最大解码轮数$n$通过在验证集上测试得到，结果表明超过3轮后不会再提升性能。尽管进一步调整$\lambda$和$\tau$可能会提高性能，但这不是本文关注的重点，因此本文未对此进行深入研究。

### 3.2 基于序列对比的源语句信息融合

在多轮解码过程中，若模型在前几轮中删除或替换了某些词语，那么这些原始词语将无法用于后续轮次的纠正，即便它们可能为后续纠正提供有价值的信息。

以图 1中的示例为例，模型应将句子：

“**我们明天从石家庄出发去北京中科院。**”

纠正为：

“**我们明天从石家庄出发去河北中国科学院。**”

在第一轮中，模型将“**中科院**”修正为“**中国科学院**”，同时将“**北京**”错误地修改为“**河北**”。在第二轮中，如果没有原始语句中的词语“**北京**”，模型就无法确定词语“**中科院**”准确的地域信息，可能会导致错误的纠错结果。“**北京**”一词的存在可以帮助模型确定原始语句中“**中科院**”一词的地域信息。

因此，在所有纠正轮次中保留完整的源语句信息可能会有利于多轮解码的性能。然而，预训练的序列到序列模型通常不具备处理源语句$x$和前一轮纠正结果$\hat{x}_{t-1}$的双编码器结构。将源语


(c) 2024 中国中文信息学会计算语言学专业委员会

句$x$和前一轮的解码结果$\hat{x}_{t-1}$直接拼接作为编码器输入会生成一个长且包含冗余信息的序列，且未改变的词语在拼接序列中会出现在两个不同的位置，不利于模型对比分析两个序列中对应位置的词语。为了有效地使用单个编码器处理源语句$x$和前一轮的解码结果$\hat{x}_{t-1}$，本文提出了一种将两者合并为单一序列的方法，如图 1所示。

具体来说，本文首先比较源语句$x$和前一轮的解码结果$\hat{x}_{t-1}$，识别出它们相同的和不同的部分。然后根据它们在原序列中的顺序将它们合并成一个序列。合并后的序列包含了未改变的词语、插入的词语和删除的词语，且保留它们的原始顺序。替换操作可以被视为插入和删除的组合。

本文通过编辑标记或特定的位置编码来区分合并序列中各词语的角色。编辑标记使用“e”（相等）、“d”（删除）和“i”（插入”）来标识合并序列中词语的状态，分别对应于同时出现在源语句$x$和前一轮的解码结果$\hat{x}_{t-1}$中的词语、仅在源语句$x$中出现的词语及新添加到上一轮解码结果$\hat{x}_{t-1}$中的词语。

本文引入一个特征向量层以学习编辑标记的表示，并在计算编码器层之前将这些编辑特征向量添加到序列到序列模型的词向量中。对于位置编码，本文在合并序列中使用了两个位置编码：源位置编码表示字符在源语句$x$中的位置，解码结果的位置编码表示其在上一轮解码结果$\hat{x}_{t-1}$中的位置。如果词语没有在某个序列中出现，则其位置编码为0。为了减少使用合并序列作为输入时新位置向量与预训练模型位置向量之间的差异，新的位置向量是基于预训练位置向量初始化的。然而，因为在使用合并序列作为输入时，位置向量会被累加两次：一次用于源位置，一次用于解码结果的位置，本文将预训练位置向量的权重减半。

## 4 实验

### 4.1 设置

为验证多轮解码方法的有效性，本文严格遵循Zhang et al. [2022a]的数据处理流程。本文使用NLPCC2018官方提供的Lang-8语料集、HSK汉语考试语料集和FCGEC训练集的组合作为训练集。NLPCC2018的训练集是从lang8收集处理得到的，是汉语的母语使用者对汉语学习者作文的修改，处理后包含1092285条数据，测试集来源于北京大学中文系构建的北京大学汉语学习者语料库，该语料库包含外国大学生的作文，旨在促进国际教育和汉语作为中介语言的研究，包含2000条数据。HSK动态作文语料库是外国汉语学习者参加等汉语水平考试等的作文答卷，处理后包含95320条数据。使用MuCGEC-dev[Zhang et al., 2022a]作为验证集，同时在NLPCC2018测试集、FCGEC的验证集[Xu et al., 2022]和NaCGEC测试集[Ma et al., 2022]上进行测试。

本文使用large设置的Transformer和预训练BART模型*作为基线，模型的训练和解码训练依据Yang and Quan [2024]的实验设定，批量大小使用1024，训练数据的最大句长使用128，最大训练周期数分别是20和10，解码的柱搜索大小是10。

为了与现有研究保持一致，对于NLPCC-2018测试数据集，本文采用官方的PKUNLP†工具对模型输出结果进行分词，并使用官方的MaxMatch（M2）‡工具计算精度（Precision，P）、召回率（Recall，R）和F0.5值。对于FCGEC验证集和NaCGEC测试集，本文使用ChERRANT进行基于字符的评估，不受分词工具对结果的影响。

### 4.2 主要结果

根据消融实验的结果，多轮解码的训练仅采用单轮解码的结果，推理采用多轮解码和早停机制（详见第 4.3节），同时使用编辑标记和位置编码来融合源语句信息（第 4.4节）。在NLPCC2018测试集的结果如表1所示。FCGEC验证集和NaCGEC测试集的结果如表2所示。

表1和2表明：1）尽管LLaMa2-7B和BaiChuan-7B大语言模型微调后的性能也较强，但其表现仍落后于经过微调的BART模型。2）多轮解码显著且一致地提升了所有基线模型的性能，相对于Transformer和BART，F0.5值分别提高了+2.78和+2.06（NLPCC2018测试集）、+6.78和+2.30（FCGEC验证集）和+4.99和+3.45（NaCGEC测试集）。3）F0.5值超过

http://lang-8.com/
*https://huggingface.co/fnlp/bart-large-chinese
†https://github.com/biug/pkunlp
‡https://github.com/nusnlp/m2scorer


(c) 2024 中国中文信息学会计算语言学专业委员会

| 模型 | P | R | F0.5 |
|---|---|---|---|
| 大模型（零样本） | | | |
| LLaMa2-7B | 11.79 | 11.46 | 11.72 |
| BaiChuan-7B | 20.87 | 23.28 | 21.31 |
| 大模型（微调） | | | |
| LLaMa2-7B | 45.85 | 27.44 | 40.43 |
| BaiChuan-7B | 51.69 | 27.92 | 44.17 |
| 序列到编辑 | | | |
| BERT-base-Chinese[Devlin et al., 2019] | 41.38 | 24.55 | 36.39 |
| HRG [Hinson et al., 2020] | 36.79 | 27.82 | 34.56 |
| SG-GEC [Wu and Wu, 2022] | 50.56 | 25.24 | 42.11 |
| 序列到序列 | | | |
| AliGM[Zhao et al., 2018] | 41.00 | 13.75 | 29.36 |
| YouDao[Fu et al., 2018] | 35.24 | 18.64 | 29.91 |
| BLCU[Li et al., 2019] | 47.63 | 12.56 | 30.57 |
| Qiu and Qu [2019] | 36.88 | 18.94 | 31.01 |
| MaskGEC[Zhao and Wang, 2020a] | 44.36 | 22.18 | 36.97 |
| GPT2-Chinese [Du, 2019] | 41.94 | 36.13 | 40.63 |
| TEA[王辰成et al., 2020] | 39.43 | 22.8 | 34.41 |
| WCDA[Tang et al., 2021] | 47.29 | 23.89 | 39.49 |
| Copy[Zhao et al., 2019] | 51.25 | 32.55 | 45.97 |
| SynGEC[Zhang et al., 2022b] | 49.96 | 33.04 | 45.32 |
| TemplateGEC [Li et al., 2023] | 54.50 | 27.40 | 45.50 |
| Alirector [Yang and Quan, 2024] | 51.76 | 33.49 | 46.67 |
| 本文 | | | |
| Transformer | 42.37 | 23.49 | 36.50 |
| + 多轮解码 | 46.64 | 24.08 | 39.28 |
| BART | 50.63 | 31.83 | 45.28 |
| + 多轮解码 | 52.56 | 33.89 | 47.34 |

表 1: 在NLPCC2018测试集上的结果

| Method | FCGEC验证集 | | | NaCGEC测试集 | | |
|---|---|---|---|---|---|---|
| | P | R | F0.5 | P | R | F0.5 |
| Transformer | 47.83 | 22.99 | 39.33 | 59.67 | 28.69 | 49.07 |
| + 多轮解码 | 58.67 | 24.76 | 46.06 | 65.61 | 31.73 | 54.06 |
| BART | 56.26 | 40.71 | 52.27 | 65.85 | 40.79 | 58.64 |
| + 多轮解码 | 59.21 | 41.57 | **54.58** | 68.82 | 44.64 | **62.09** |

表 2: 在FCGEC验证集和NaCGEC测试集上的结果

CCL 2024

| 设置 | FCGEC验证集 | | | NaCGEC测试集 | | | Speed |
|---|---|---|---|---|---|---|---|
| | P | R | F0.5 | P | R | F0.5 | |
| BART | 56.26 | 40.71 | 52.27 | 65.85 | 40.79 | 58.64 | 1.00x |
| 训练 | | | | | | | |
| 单次 | 59.21 | 41.57 | **54.58** | 68.82 | 44.64 | **62.09** | 0.83x |
| 多轮 | 57.86 | 42.45 | 53.94 | 67.51 | 43.97 | 60.98 | |
| 解码 | | | | | | | |
| 策略模型[2018] | 58.07 | 40.09 | 53.29 | 67.22 | 43.52 | 60.62 | 0.27x |
| 无早停机制 | | | | | | | |
| $n = 1$ | 58.47 | 40.91 | 53.85 | 67.66 | 43.42 | 60.86 | 0.46x |
| $n = 2$ | 58.13 | 41.09 | 53.68 | 67.24 | 43.48 | 60.62 | 0.41x |
| $n = 3$ | 58.18 | 41.21 | 53.75 | 67.27 | 43.68 | 60.71 | 0.38x |
| 早停机制, $n = 3$ | 59.21 | 41.57 | **54.58** | 68.82 | 44.64 | **62.09** | 0.83x |

表 3: 各种多轮解码训练和推理设置的结果。速度是在NaCGEC测试集上的推理速度。

了最新的Alirector模型，提高了+0.67。尽管本文仅在广泛使用的Transformer和BART基线上应用了本文的方法，但本文认为本文的方法也可以用于更先进的基线模型，并带来更好的性能。

### 4.3 多轮解码训练和推理的消融实验

除了训练模型生成给定输入源语句$x$的参考答案$r$之外，多轮解码训练还可以进行单轮解码，如Omelianchuk et al. [2020]所述，或者进行几轮解码直至满足推理终止条件。本文探讨了在推理时采用多轮解码并带有早停机制的多轮解码训练中，单次解码与多轮解码的效果。其中，“多轮解码”是指解码方法，模型训练需要使用之前的解码结果，此结果可由单次解码产生，也可由多轮解码产生。

对于单次多轮解码训练，本文使用模型将输入源语句$x$解码为单次解码结果$\hat{x}_0$，并训练模型在给定源语句$x$和单次解码结果$\hat{x}_0$的情况下生成参考答案$r$：

$$M(x, \hat{x}_0) \rightarrow r \tag{7}$$

对于多遍多轮解码训练，本文以源语句$x$作为初始解码结果$\hat{x}_{-1}$开始，迭代地将前一轮解码结果$\hat{x}_{i-1}$解码为本轮解码结果$\hat{x}_i$，重复数轮直至满足终止条件，并训练模型在给定源语句$x$和本轮解码结果$\hat{x}_i$的情况下生成参考答案$r$：

$$M(x, \hat{x}_i) \rightarrow r \tag{8}$$

本文还探讨了在多轮解码推理中最大解码轮数，以及是否使用早停机制的的影响，采用单次多轮解码进行训练。此外，本文对比了本文提出的早停机制与Geng et al. [2018]提出的策略模型。Geng et al. [2018]采用强化学习方法，通过评估连续两次解码遍之间的差异来确定解码轮数，并针对机器翻译任务使用BLEU作为奖励优化策略模型。在本文的语法纠错任务实验中，奖励基于F0.5得分而非BLEU，结果如表3所示。其中，“训练”模块中的多轮是指模型在训练时使用带有早停机制的多轮解码算法产生训练所需的上一轮解码结果，具体的轮数受早停机制影响是变化的。

表3表明：1) 对于多轮解码训练，两种设置均能带来性能提升，但单次解码设置在训练效率更高的情况下,相较于多轮解码，在FCGEC验证集和NaCGEC测试集上都取得了更高的F0.5值（+0.64和+1.11）。2) 不同的最大解码轮数取得了相近的F0.5值，更大的最大解码轮数会导致推理速度下降，但早停机制可以有效缓解这一问题，并取得了最佳的性能，3) 基于策略网络的多轮解码在两个共享任务上同样带来了F0.5的改进，但本文简单的早停方法比策略网络[Geng et al., 2018]更有效率，并在FCGEC验证集和NaCGEC测试集上都取得了更高的F0.5值（+1.29，+1.47）。

CCL 2024

| 方法 | FCGEC验证集 | | | NaCGEC测试集 | | |
|---|---|---|---|---|---|---|
| | P | R | F0.5 | P | R | F0.5 |
| BART | 56.26 | 40.71 | 52.27 | 65.85 | 40.79 | 58.64 |
| None | 56.72 | 43.11 | 53.35 | 65.64 | 46.13 | 60.52 |
| Concat | 57.32 | 42.15 | 53.47 | 66.90 | 44.11 | 60.63 |
| Edit | 57.59 | 42.97 | 53.92 | 67.02 | 44.34 | 60.80 |
| Pos | 57.86 | 42.81 | 54.06 | 67.41 | 44.00 | 60.93 |
| Pos+edit | 59.21 | 41.57 | **54.58** | 68.82 | 44.64 | **62.09** |

表 4: 不同源语句信息融合方法的结果

现有最好的机器翻译多轮解码研究[Geng et al., 2018]采用了复杂的强化学习方法来决定所需的解码轮数。然而，强化学习训练可能不稳定，导致性能波动。本文的监督方法直接基于解码序列的向量表示训练简单的二分类器，本文认为本文的早停方法易于实现，并且在实践中证明了有效性。

### 4.4 源语句信息融合的影响分析

本文评估了不同源语句信息融合方法对BART的影响，包括：1）在多轮解码推理中仅使用前一轮的解码结果$\hat{x}_{t-1}$而非同时使用前一轮的解码结果$\hat{x}_{t-1}$和源语句$x$（“None”），2）序列拼接（“Concat”），3）编辑标记（“Edit”），4）位置编码（“Pos”），以及5）编辑标记和位置编码的组合（“Pos+Edit”）。实验结果如表4所示。

表4表明：1）未融合源语句信息的普通多轮解码（“None”）在FCGEC验证集和NaCGEC测试集上分别带来了+1.08和+1.88的F0.5值提升，证明了多轮解码的有效性，2）通过序列拼接（“Concat”）提供源语句信息较没有源语句信息（“None”）可以在FCGEC验证集和NaCGEC测试集上分别带来了+0.12和+0.11的F0.5值提升，证明了源语句信息的正面影响，3）位置编码（“Pos”）和编辑标签（“Edit”）都显著优于序列拼接，表明基于序列对比的源语句信息融合方法的重要性，其中位置编码的性能略优于编辑标签，这可能是因为预训练位置向量的初始化，4）位置编码与编辑标记的组合（“Pos+Edit”）带来了最佳性能。

## 5 相关研究

**用于语法纠错的序列到编辑方法** 可将语法纠错任务视为序列标注任务，通过迭代预测句子中每个位置的编辑标记。受GECToR[Omelianchuk et al., 2020]启发，Liang et al. [2020]将中文语法纠错任务视作序列标注任务。Zhang et al. [2022a]直接采用GECToR，并利用预训练模型进行增强。TtT[Li and Shi, 2021]提出了一种非自回归语法纠错方法，结合了BERT[Devlin et al., 2019]编码器和CRF编码器[Lafferty et al., 2001]。Li et al. [2022b]则提出一个序列到编辑模型，用于中文语法纠错任务，在每个字符前预测一个动作，动作由跳过、复制和生成三类构成，然后将这些动作与基本序列到序列框架融合，提供最终预测。Tan et al. [2023]提出了一种基于纠错类型的校正方法，在每个训练实例中，额外构建多个涉及特定类型错误训练实例，然后使用这些额外构造的训练实例和原始的训练实例依次训练模型。序列到编辑模型的一个局限性在于其对编辑操作手动构建的依赖，这降低了模型的可迁移性并可能影响输出的流畅性[Li et al., 2023]。上述方法均使用了多轮解码。

**用于语法纠错的序列到序列方法** 可将语法纠错视为序列生成任务[Fang et al., 2023, Li et al., 2022a, Liu et al., 2021b, Wang et al., 2021]。早期研究主要采用基于CNN[Ren et al., 2018]或RNN[Zhou et al., 2018]的模型来解决语法纠错任务。随着transformer[Vaswani et al., 2017]在机器翻译任务中取得巨大成功，它也被应用于语法纠错任务。HRG[Malmi et al., 2019]结合了拼写检查器、NMT-base模型和序列编辑模型，但其拼写检查器基于语言模型，未能充分利用上下文。Kaneko et al. [2020]首先对MLM模型进行微调，然后将微调后的模型输出作为额外特征用于语法纠错模型，以最大化MLM模型的优势。Zhao and Wang [2020b]提出了数据增强方法MaskGEC，该方法在训练过程中动态地对输入句子添加随机噪声。Tang et al. [2021]提出了使用数据增强的方法来扩展训练数据，以提升模型性能。Zhou et al. [2020]通过编


(c) 2024 中国中文信息学会计算语言学专业委员会

辑式投票机制集成了序列到序列模型和序列到编辑模型。Zhang et al. [2022a]提出了基于序列到序列架构的中文BART模型，适用于中文语法纠错任务。Fei et al. [2023], Zhang et al. [2022b]通过添加额外的语法知识或利用特定语言知识来提高纠错性能。Sun and Wang [2022]提出通过权衡精确度-召回率来提高纠错性能。Zhang et al. [2023]提出通过重排序来选择正确率更高的纠错结果，从而提高纠错性能。

**多轮解码在NMT中的应用** 多轮解码已被用于改进神经机器翻译（NMT）[Xia et al., 2017, Mahmood et al., 2017, Zhang et al., 2018, Geng et al., 2018, Liu et al., 2019]。自动后编辑（APE）也可视为多轮解码的一种特例[Correia and Martins, 2019, Pal et al., 2020, Bhattacharyya et al., 2022, Jung et al., 2023]。这些研究均强调了源信息融合对性能提升的重要性，但它们在源输入和已解码序列上采用了双编码器结构，这主要是因为它们采用不同的语言，这两种语言在拼写和结构上的差异。然而，在语法纠错任务中，源句子和解码序列通常较为接近。本文首次：1）研究了中文语法纠错任务的序列到序列模型多轮解码，2）提出了一种简单有效的早停机制来解决效率问题，3）提出了一种将源语句和解码序列合并的方法，以适应只有一个编码器的预训练序列到序列模型。

## 6 总结

本文针对现有的序列到序列纠错方法仅进行一次从左到右的解码，在解码过程中不能考虑后续词语的问题，提出通过多轮解码来重复检查纠正现有的纠错结果，改进序列到序列语法纠错方法。为提升多轮解码的效率，本文提出早停机制，在下一轮可能不会产生更好的纠错结果时，提前结束多轮解码。本文提出源信息融合方法来解决多轮解码中删除和替换词语导致的信息丢失问题，并根据两个序列相近的特性，提出基于对比的序列混合方法，避免序列拼接导致的效率问题。

在NLPCC2018测试集、FCGEC验证集和NaCGEC测试集上的实验表明，本文的方法相对于强大的BART-large基线带来了+2.06，+2.31和+3.45的F0.5值提升，分别取得了47.34，54.58和62.09的F0.5值，证明了本文方法的有效性。

## 致谢

感谢审稿人帮助改善论文的建议。本文受国家自然科学基金青年项目（62306284），中国博士后科学基金面上项目（2023M743189），河南省自然科学基金青年项目（232300421386）资助。

## References

Pushpak Bhattacharyya, Rajen Chatterjee, Markus Freitag, Diptesh Kanojia, Matteo Negri, and Marco Turchi. Findings of the wmt 2022 shared task on automatic post-editing. In *Proceedings of the Seventh Conference on Machine Translation*, pages 109–117, Abu Dhabi, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.5.

Gonçalo M. Correia and André F. T. Martins. A simple and effective approach to automatic post-editing with transfer learning. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3050–3056, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1292. URL https://aclanthology.org/P19-1292.

Michael J. Denkowski. Machine translation for human translators. 2015. URL https://api.semanticscholar.org/CorpusID:8279024.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.


(c) 2024 中国中文信息学会计算语言学专业委员会

Zeyao Du. Gpt2-chinese: Tools for training gpt2 model in chinese language. https://github.com/Morizeyao/GPT2-Chinese, 2019.

Tao Fang, Jinpeng Hu, Derek F. Wong, Xiang Wan, Lidia S. Chao, and Tsung-Hui Chang. Improving grammatical error correction with multimodal feature integration. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9328–9344, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.594. URL https://aclanthology.org/2023.findings-acl.594.

Yuejiao Fei, Leyang Cui, Sen Yang, Wai Lam, Zhenzhong Lan, and Shuming Shi. Enhancing grammatical error correction systems with explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7489–7501, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.413. URL https://aclanthology.org/2023.acl-long.413.

Kai Fu, Jin Huang, and Yitao Duan. Youdao's winning solution to the nlpcc-2018 task 2 challenge: A neural machine translation approach to chinese grammatical error correction. In *Natural Language Processing and Chinese Computing*, 2018. URL https://api.semanticscholar.org/CorpusID:52007443.

Xinwei Geng, Xiaocheng Feng, Bing Qin, and Ting Liu. Adaptive multi-pass decoder for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 523–532, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1048. URL https://aclanthology.org/D18-1048.

Charles Hinson, Hen-Hsen Huang, and Hsin-Hsi Chen. Heterogeneous recycle generation for Chinese grammatical error correction. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2191–2201, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.199. URL https://aclanthology.org/2020.coling-main.199.

Baikjin Jung, Myungji Lee, Jong-Hyeok Lee, and Yunsu Kim. Bring more attention to syntactic symmetry for automatic postediting of high-quality machine translations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1433–1441, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.122. URL https://aclanthology.org/2023.acl-short.122.

Masahiro Kaneko, Masato Mita, Shun Kiyono, Jun Suzuki, and Kentaro Inui. Encoder-decoder models can benefit from pre-trained masked language models in grammatical error correction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4248–4254, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.391. URL https://aclanthology.org/2020.acl-main.391.

Shun Kiyono, Jun Suzuki, Masato Mita, Tomoya Mizumoto, and Kentaro Inui. An empirical study of incorporating pseudo data into grammatical error correction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1236–1242, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1119. URL https://aclanthology.org/D19-1119.

John D. Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning*, 2001. URL https://api.semanticscholar.org/CorpusID:219683473.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-


(c) 2024 中国中文信息学会计算语言学专业委员会

training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL https://aclanthology.org/2020.acl-main.703.

Bei Li, Quan Du, Tao Zhou, Yi Jing, Shuhan Zhou, Xin Zeng, Tong Xiao, JingBo Zhu, Xuebo Liu, and Min Zhang. ODE transformer: An ordinary differential equation-inspired model for sequence generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8335–8351, Dublin, Ireland, May 2022a. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.571. URL https://aclanthology.org/2022.acl-long.571.

Jiquan Li, Junliang Guo, Yongxin Zhu, Xin Sheng, Deqiang Jiang, Bo Ren, and Linli Xu. Sequence-to-action: Grammatical error correction with action guided sequence generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10974–10982, Jun. 2022b. doi: 10.1609/aaai.v36i10.21345. URL https://ojs.aaai.org/index.php/AAAI/article/view/21345.

Piji Li and Shuming Shi. Tail-to-tail non-autoregressive sequence prediction for Chinese grammatical error correction. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4973–4984, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.385. URL https://aclanthology.org/2021.acl-long.385.

Si Li, Jianbo Zhao, Guirong Shi, Yuanpeng Tan, Huifang Xu, Guang Chen, Haibo Lan, and Zhiqing Lin. Chinese grammatical error correction based on convolutional sequence to sequence model. volume 7, pages 72905–72913, 2019. doi: 10.1109/ACCESS.2019.2917631.

Yinghao Li, Xuebo Liu, Shuo Wang, Peiyuan Gong, Derek F. Wong, Yang Gao, Heyan Huang, and Min Zhang. TemplateGEC: Improving grammatical error correction with detection template. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6878–6892, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.380. URL https://aclanthology.org/2023.acl-long.380.

Deng Liang, Chen Zheng, Lei Guo, Xin Cui, Xiuzhang Xiong, Hengqiao Rong, and Jinpeng Dong. BERT enhanced neural machine translation and sequence tagging model for Chinese grammatical error diagnosis. In Erhong YANG, Endong XUN, Baolin ZHANG, and Gaoqi RAO, editors, *Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 57–66, Suzhou, China, December 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.nlptea-1.8.

Cao Liu, Shizhu He, Kang Liu, and Jun Zhao. Vocabulary pyramid network: Multi-pass encoding and decoding with multi-level vocabularies for response generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3774–3783, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1367. URL https://aclanthology.org/P19-1367.

Xuebo Liu, Longyue Wang, Derek F. Wong, Liang Ding, Lidia S. Chao, and Zhaopeng Tu. Understanding and improving encoder layer fusion in sequence-to-sequence learning. 2021a. URL https://openreview.net/forum?id=n1HD8M6WGn.

Xuebo Liu, Longyue Wang, Derek F. Wong, Liang Ding, Lidia S. Chao, and Zhaopeng Tu. Understanding and improving encoder layer fusion in sequence-to-sequence learning. 2021b. URL https://openreview.net/forum?id=n1HD8M6WGn.

Shirong Ma, Yinghui Li, Rongyi Sun, Qingyu Zhou, Shulin Huang, Ding Zhang, Li Yangning, Ruiyang Liu, Zhongli Li, Yunbo Cao, Haitao Zheng, and Ying Shen. Linguistic rules-based


(c) 2024 中国中文信息学会计算语言学专业委员会

corpus generation for native Chinese grammatical error correction. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 576–589, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.40. URL https://aclanthology.org/2022.findings-emnlp.40.

Rehan Mahmood, Zulin Wang, and Qin Huang. Multi-pass decoding for the robust transmission of deep-space images. In *2017 IEEE 85th Vehicular Technology Conference (VTC Spring)*, pages 1–5, 2017. doi: 10.1109/VTCSpring.2017.8108437.

Eric Malmi, Sebastian Krause, Sascha Rothe, Daniil Mirylenka, and Aliaksei Severyn. Encode, tag, realize: High-precision text editing. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5054–5065, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1510. URL https://aclanthology.org/D19-1510.

Hwee Tou Ng, Siew Mei Wu, Yuanbin Wu, Christian Hadiwinoto, and Joel Tetreault. The CoNLL-2013 shared task on grammatical error correction. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–12, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://aclanthology.org/W13-3601.

Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. The CoNLL-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–14, Baltimore, Maryland, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-1701. URL https://aclanthology.org/W14-1701.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. GECToR – grammatical error correction: Tag, not rewrite. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–170, Seattle, WA, USA → Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.bea-1.16. URL https://aclanthology.org/2020.bea-1.16.

Santanu Pal, Hongfei Xu, Nico Herbig, Sudip Kumar Naskar, Antonio Krüger, and Josef van Genabith. The transference architecture for automatic post-editing. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5963–5974, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.524. URL https://aclanthology.org/2020.coling-main.524.

Zhaoquan Qiu and Youli Qu. A two-stage model for chinese grammatical error correction. *IEEE Access*, 7:146772–146777, 2019. doi: 10.1109/ACCESS.2019.2940607.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1), jan 2020. ISSN 1532-4435.

Hongkai Ren, Liner Yang, and Endong Xun. A sequence to sequence learning for chinese grammatical error correction. In Min Zhang, Vincent Ng, Dongyan Zhao, Sujian Li, and Hongying Zan, editors, *Natural Language Processing and Chinese Computing*, pages 401–410, Cham, 2018. Springer International Publishing.

Sascha Rothe, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. A simple recipe for multilingual grammatical error correction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 702–707, Online,

August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.89. URL https://aclanthology.org/2021.acl-short.89.

Felix Stahlberg and Shankar Kumar. Seq2Edits: Sequence transduction using span-level edit operations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5147–5159, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.418. URL https://aclanthology.org/2020.emnlp-main.418.

Xin Sun and Houfeng Wang. Adjusting the precision-recall trade-off with align-and-predict decoding for grammatical error correction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 686–693, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-short.77. URL https://aclanthology.org/2022.acl-short.77.

Xin Sun, Tao Ge, Furu Wei, and Houfeng Wang. Instantaneous grammatical error correction with shallow aggressive decoding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5937–5947, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.462. URL https://aclanthology.org/2021.acl-long.462.

Minghuan Tan, Min Yang, and Ruifeng Xu. Focal training and tagger decouple for grammatical error correction. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5978–5985, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.370. URL https://aclanthology.org/2023.findings-acl.370.

Zecheng Tang, Yixin Ji, Yibo Zhao, and Junhui Li. 基于字词粒度噪声数据增强的中文语法纠错(Chinese grammatical error correction enhanced by data augmentation from word and character levels). In Sheng Li, Maosong Sun, Yang Liu, Hua Wu, Kang Liu, Wanxiang Che, Shizhu He, and Gaoqi Rao, editors, *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 813–824, Huhhot, China, August 2021. Chinese Information Processing Society of China. URL https://aclanthology.org/2021.ccl-1.73.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Yu Wang, Yuelin Wang, Kai Dang, Jie Liu, and Zhuo Liu. A comprehensive survey of grammatical error correction. *ACM Trans. Intell. Syst. Technol.*, 12(5), dec 2021. ISSN 2157-6904. doi: 10.1145/3474840. URL https://doi.org/10.1145/3474840.

Xiuyu Wu and Yunfang Wu. From spelling to grammar: A new framework for Chinese grammatical error correction. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 889–902, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.63. URL https://aclanthology.org/2022.findings-emnlp.63.

Yingce Xia, Fei Tian, Lijun Wu, Jianxin Lin, Tao Qin, Nenghai Yu, and Tie-Yan Liu. Deliberation networks: Sequence generation beyond one-pass decoding. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/c6036a69be21cb660499b75718a3ef24-Paper.pdf.


(c) 2024 中国中文信息学会计算语言学专业委员会

Lvxiaowei Xu, Jianwang Wu, Jiawei Peng, Jiayu Fu, and Ming Cai. FCGEC: Fine-grained corpus for Chinese grammatical error correction. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1900–1918, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.137. URL https://aclanthology.org/2022.findings-emnlp.137.

Haihui Yang and Xiaojun Quan. Alirector: Alignment-enhanced chinese grammatical error corrector. *ArXiv*, abs/2402.04601, 2024. URL https://arxiv.org/abs/2402.04601.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf.

Zheng Yuan, Shiva Taslimipoor, Christopher Davis, and Christopher Bryant. Multi-class grammatical error detection for correction: A tale of two systems. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8722–8736, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.687. URL https://aclanthology.org/2021.emnlp-main.687.

Xiangwen Zhang, Jinsong Su, Yue Qin, Yang Liu, Rongrong Ji, and Hongji Wang. Asynchronous bidirectional decoding for neural machine translation. AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018. ISBN 978-1-57735-800-8.

Ying Zhang, Hidetaka Kamigaito, and Manabu Okumura. Bidirectional transformer reranker for grammatical error correction. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3801–3825, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.234. URL https://aclanthology.org/2023.findings-acl.234.

Yue Zhang, Zhenghua Li, Zuyi Bao, Jiacheng Li, Bo Zhang, Chen Li, Fei Huang, and Min Zhang. MuCGEC: a multi-reference multi-source evaluation dataset for Chinese grammatical error correction. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3118–3130, Seattle, United States, July 2022a. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.227. URL https://aclanthology.org/2022.naacl-main.227.

Yue Zhang, Bo Zhang, Zhenghua Li, Zuyi Bao, Chen Li, and Min Zhang. SynGEC: Syntax-enhanced grammatical error correction with a tailored GEC-oriented parser. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2518–2531, Abu Dhabi, United Arab Emirates, December 2022b. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.162. URL https://aclanthology.org/2022.emnlp-main.162.

Wei Zhao, Liang Wang, Kewei Shen, Ruoyu Jia, and Jingming Liu. Improving grammatical error correction via pre-training a copy-augmented architecture with unlabeled data. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 156–165, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1014. URL https://aclanthology.org/N19-1014.

Yuanyuan Zhao, Nan Jiang, Weiwei Sun, and Xiaojun Wan. Overview of the NLPCC 2018 shared task: Grammatical error correction. In Min Zhang, Vincent Ng, Dongyan Zhao, Sujian Li, and


(c) 2024 中国中文信息学会计算语言学专业委员会

Hongying Zan, editors, *Natural Language Processing and Chinese Computing - 7th CCF International Conference, NLPCC 2018, Hohhot, China, August 26-30, 2018, Proceedings, Part II*, volume 11109 of *Lecture Notes in Computer Science*, pages 439–445. Springer, 2018. doi: 10.1007/978-3-319-99501-4_41. URL https://doi.org/10.1007/978-3-319-99501-4_41.

Zewei Zhao and Houfeng Wang. Maskgec: Improving neural grammatical error correction via dynamic masking. In *AAAI Conference on Artificial Intelligence*, 2020a. URL https://api.semanticscholar.org/CorpusID:213955573.

Zewei Zhao and Houfeng Wang. Maskgec: Improving neural grammatical error correction via dynamic masking. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):1226–1233, Apr. 2020b. doi: 10.1609/aaai.v34i01.5476. URL https://ojs.aaai.org/index.php/AAAI/article/view/5476.

Junpei Zhou, Chen Li, Hengyou Liu, Zuyi Bao, Guangwei Xu, and Linlin Li. *Chinese Grammatical Error Correction Using Statistical and Neural Models: 7th CCF International Conference, NLPCC 2018, Hohhot, China, August 26–30, 2018, Proceedings, Part II*, pages 117–128. 08 2018. ISBN 978-3-319-99500-7. doi: 10.1007/978-3-319-99501-4_10.

Wangchunshu Zhou, Tao Ge, Chang Mu, Ke Xu, Furu Wei, and Ming Zhou. Improving grammatical error correction with machine translation pairs. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 318–328, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.30. URL https://aclanthology.org/2020.findings-emnlp.30.

王辰成, 杨麟儿, 王莹莹, 杜永萍, and 杨尔弘. 基于transformer增强架构的中文语法纠错方法. 中文信息学报, 34(6):106, 2020. URL http://jcip.cipsc.org.cn/CN/Y2020/V34/I6/106.

CCL 2024

# 中西谚语多元价值观资源库建设及对比研究

**杜霞[1], 刘鹏远[1,2,*]，于东[1]**
1.北京语言大学信息科学学院，北京，100083
2.国家语言资源监测与研究平面媒体中心，北京，100083
202321198091@stu.blcu.edu.cn
liupengyuan@pku.edu.cn，yudong@blcu.edu.cn

## 摘要

中西方谚语是中西方文化的结晶，分别蕴含着中西方文化中最基本的价值观。但目前缺乏中西方谚语价值观资源，难以对谚语所体现的中西方价值观进行全面的研究，特别是定量对比研究。因此本文设计了多元价值观体系，包含动机及需求、共同及特色价值观、价值判断和使用场景，根据这个体系构建了中西方谚语多元价值观资源库并进行了考察与对比分析。本文发现中西谚语在价值判断、使用场景及部分价值观上具有相似性，在具体内涵表达上各具独特性。

**关键词：** 中西方谚语 ； 价值观 ； 跨文化对比

# The construction and comparative study of the resource library of Chinese and Western proverbs and multiple values

**Xia Du[1], Pengyuan Liu[1,2,*],Dong Yu[1]**
1.Faculty of Computer Science,Beijing Language and Culture University,Beijing,100083
2.National Language Resources Monitoring and Research Center for Print Media,Beijing,100083
202321198091@stu.blcu.edu.cn
liupengyuan@pku.edu.cn，yudong@blcu.edu.cn

## Abstract

Chinese and Western proverbs are the crystallization of Chinese and Western culture, and contain the most basic values of Chinese and Western culture respectively. However, due to the lack of Chinese and western values resources, it is difficult to conduct a comprehensive study on the Chinese and western values embodied in proverbs, especially a quantitative comparative study. Therefore, this paper designs a multi-value system, including motivation and needs, common and characteristic values, value judgments and usage scenarios. Based on this system, a multi-value resource database of Chinese and Western proverbs is constructed and investigated and analyzed. This paper finds that Chinese and Western proverbs are similar in value judgment, usage scenarios and some values, and unique in specific connotation expression.

**Keywords:** Chinese and western proverbs , Values , Cross-cultural comparison

*为通讯作者
**基金项目**：中央高校基本科研业务费(北京语言大学梧桐创新平台，21PT04)
©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

## 1 引言

价值观是人区分好坏、美丑、益损、正确与错误，符合或违背自己意愿等的观念系统，它通常是充满情感的，并为人的正当行为提供充分的理由(黄希庭et al., 1994)。由于环境、历史等因素的影响，中西价值观念形成了各自不同的风格。通过分析中西方价值观的异同，有助于深入理解一个文化群体的心理和行为模式，理解传统与现代之间的关系，并且在多元文化交流中，可以减少文化冲突和误解，促进文化交流和融合。

谚语是人的实际经验之结果，而用美的言辞以表达者，于日常谈话可以公然使用，而规定人的行为之言语(郭绍虞, 1925)。谚语作为人们喜闻乐见的语言形式，深受宗教、文学、伦理道德等社会意识形态的影响，留下了鲜明的文化烙印(王德春, 2003)，蕴含了一个民族最传统的价值观和思维方式。近年来，许多学者致力于构建反映价值观的数据资源库，但是现有研究主要聚焦于句子级别的分析，利用自然语言处理（NLP）技术从大量文本数据中提取价值观信息，并着重于探究语言如何传递价值观，以及这些价值观是如何与个体的行为决策相联系的(Ishita et al., 2010; Liu et al., 2022)。基于谚语价值观的研究相对较少，且数据量有限，通常仅以少数谚语作为例证进行定性分析(朱益平, 2009)，而且这种选择性的样本无法较为全面地反映谚语在不同文化中的价值观。在中西方谚语的比较研究方面，已有学者对谚语背后的文化、意识形态以及社会价值观进行了深入探讨，但这些研究往往聚焦于揭示中西方文化在价值观念、思维模式和社会习俗方面的差异性(侯洁诚, 2017)，缺乏从多元价值角度对谚语进行对比分析，无法全面地展现和反映一个民族的传统与文化。

基于此，本文搜集了中西方谚语共2278条，标注了动机及需求层次、多元价值观、价值判断和使用场景。然后，对中西方谚语中蕴含的价值观、动机和需求层次等进行了量化分析，并研究了各个价值在施瓦茨动机层次与马斯洛需求层次上的分布情况以及中西谚语在词汇层面的价值表达。发现中西谚语在价值判断、使用场景及表达“公平”、“正义”、“友善”等价值观时具有相似性，但在具体内涵表达上又各具独特性。

本文贡献主要包括以下三个方面：

（1）在资源建设方面，设计了多元价值观资源体系，综合施瓦茨动机和马斯洛需求层次、全人类共同价值观、中国传统价值观、中国社会主义核心价值观、价值倾向判断与使用场景，搜集并整理了2278条中西方谚语，构建了首个具有多元价值的中西方谚语资源库；

（2）在文化研究方面，首次对谚语进行了动机和需求层次分析，深入挖掘了谚语背后的心理机制与人类需求，为传统文化研究提供了新视角；

（3）在跨文化对比方面，基于多元价值中西方谚语资源库，首次以定量分析视角对中西方谚语的价值观、动机、需求进行跨文化对比，揭示了不同文化背景下的价值观、动机及需求表达的相似性与差异性。

## 2 相关工作

### 2.1 中西谚语及其比较研究

谚语研究作为一个传统的课题，历来是语言学、社会学等多学科领域研究的重要问题。目前关于谚语的研究涉及范围极广，包括语用标记倾向、话语功能、语义认知分析等(马利军and 张积家, 2017)。但关于中西谚语的比较研究相对较少，尤其是在研究中西谚语所反映的价值观方面，相关研究更是不足。朱益平(朱益平, 2009)采用克拉克洪的文化模式分析方法，通过对比众多汉英谚语，探讨了中西文化价值观的不同。侯洁诚(侯洁诚, 2017)从地理环境、价值观念、宗教信仰、社会历史和文学艺术等方面对英汉谚语进行了对比分析。钟鸣(钟鸣and 代玮炜, 2008)则指出生存环境和地理区域、习俗、宗教和历史典故和寓言的差异折射出中西方的文化差异，并从个人主义与集体主义、人与自然的关系、性恶论与性善论、等级观念等角度对中西方的意识形态进行了深入讨论。但这些研究多为选择部分实例进行定性分析，缺乏利用全面系统的资源库对谚语进行定量研究。

### 2.2 价值观资源建设研究

在科技进步的推动下，以人工智能为核心的计算机科学已经应用到多个学科领域，近年来许多学者开始利用一些自然语言处理技术构建有关价值观方面的数据资源。

Jilin chen等(Chen et al., 2014)首次分析了基本人类价值观与他们在在线社交媒体中的用词之间的联系。Boyd等(Boyd et al., 2015)使用了一个包含着人类语言、价值观和行为的数据


(c) 2024 中国中文信息学会计算语言学专业委员会

集，研究开放式文本中出现的价值观和行为，然后将研究结果应用于Facebook 用户状态更新的语料库中，来探究核心价值观如何影响用户通过社交媒体分享个人的想法和行为。Liang Qiu等(Qiu et al., 2022)构建了一个包含21374个文本场景的人类价值观数据集。Emi Ishita 等(Ishita et al., 2010)研究了人类价值观的多标签分类，选取了2005个句子进行人工标注和机器分类，以识别表达人类价值观的文本，发现人工标注具有较好的一致性，几个分类器的表现也超出了预期。刘鹏远等(Liu et al., 2022)基于核心价值观建立了首个面向价值观计算的中文核心价值-行为体系及相应的知识库，并提出了3个价值观计算任务且进行了实验，实验结果表明基于预训练语言模型的方法在价值观方向判别上表现优异。

总体来看，目前缺少对大规模中西谚语进行多元价值观分析的研究，而且关于价值观的研究资料大部分以英文为主，中文领域缺乏蕴含价值观知识的大规模数据资源。因此，本文对中西方谚语进行价值观计算与分析，希望能够补充价值观资源建设，并更好地理解文化多样性。

## 3 多维价值观体系设计

### 3.1 框架设计

价值与人类的基本需求和动机密不可分。Schwartz认为价值观起源于人类的基本动机，人类具有一些动机目标，最终的价值观正是由这一组具体化的动机目标构成，而这些动机目标源自人类的基本需求(Schwartz and Bilsky, 1987)。因此，本文将动机和需求作为价值体系的两个基础层。

在基础层之上，设置了共同价值层。共同价值层为习近平主席出席第七十届联合国大会时提出的全人类共同价值。全人类共同价值观着眼于全人类的历史发展进程，是超越国家、民族和文化界限的普世价值。在共同价值层之上，设置了特色价值层，将包含中西方各种特色文化的价值。在特色价值层之上，设置了应用层，将包含对谚语的价值正负向判断与使用场景细分。

最终，多维价值观体系的框架设计如图1所示。

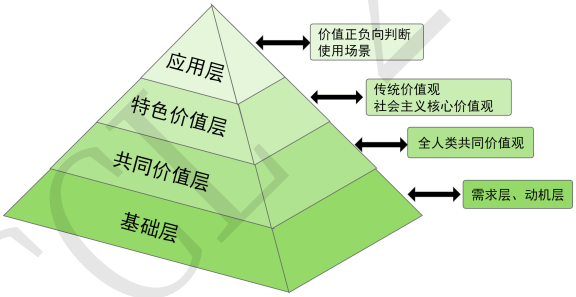



Figure 1: 多维价值观体系框架

### 3.2 具体价值观选取

**需求层。**马斯洛需求层次理论作为心理学领域的经典理论之一，被广泛地应用于各种研究领域，该理论将人类需求划分为五个层次，依次为生理需求、安全需求、爱和归属感需求、尊重需求以及自我实现需求，这五种需求是人类生存和发展所需要的基本需求(Maslow, 1943)。

**动机层。**Schwartz提出，人类生存的“普遍”的动机目标主要有三类：一是个体的动机；二是协调社会协作的动机；三是维护群体生存和利益而产生的动机。最终确定了安全、自我导向、刺激、传统、成就、遵从、权力、慈善、大同主义和享乐主义这十种人类基础动机类型。

**共同价值层。**2015年9月28日，习近平主席出席第七十届联合国大会讲话指出：“和平、发展、公平、正义、民主、自由是全人类的共同价值，也是联合国的崇高目标”。

**特色价值层**[1] **—中国传统价值观。**价值观念的形成深受不同时代和社会环境的影响，呈现

[1]在特色价值层，因西方文化的各类特色价值观体系并不凸显，多以施瓦茨动机为价值观框架，具体的价值观需要对各国进行大规模各语种的统计进一步分析构建，本文是对多维价值观体系进行的初步尝试，当前本文仅选取了中国传统价值观与中国社会主义核心价值观，未来工作中将继续丰富特色价值体系，扩大中西方价值观的比较范


(c) 2024 中国中文信息学会计算语言学专业委员会

出多元且复杂的特性。中华文化源远流长，从古至今，价值观也在不断变化。但传统价值观是一个民族长期积累的精神财富，它承载了历史、文化和道德的精髓。戴木才(2015)在《中国人的美德与核心价值观》中提出“仁义礼智信”是中华民族的传统核心价值观，但“仁义礼智信”并不能完全覆盖传统美德德目的丰富内容，忠、孝、廉、耻、勇这些传统美德德目，在我国历史发展进程中，一直发挥着重要的规范作用。

**特色价值层—中国社会主义核心价值观。**随着时代的更迭，价值观也在不断发展。党的十八大提出：倡导富强、民主、文明、和谐，倡导自由、平等、公正、法治，倡导爱国、敬业、诚信、友善，积极培育和践行社会主义核心价值观。富强、民主、文明、和谐是国家层面的价值目标，自由、平等、公正、法治是社会层面的价值目标，爱国、敬业、诚信、友善是公民个人层面的价值准则。

**应用层—价值正负向判断。**人们对各种社会现象、问题，往往会作出好与坏或应该与否的判断(W.D.拉蒙特, 1992)。本文从价值判断角度将谚语表达的价值观进一步细分为正向和负向两个类别。

**应用层—使用场景。**根据谚语所体现的价值观所使用的场景，本文将其划分为国家、社会、个人以及家庭四个部分。

表1是本文的多维价值观体系与具体价值观。可以看出，由于中西方文化之间价值观存在一定程度上的共通性，若干价值观如：仁、义、礼、智、信、自由、平等、民主等在多个层次中均有体现。但该体系的动机是将从不同层面和特色价值观进行中西方价值观的宏观对比分析，并不仅仅从单个价值观视角，此外，层次之间并不是完全割裂的状态，特色价值层建立在共同价值层这个基础框架之上，因此允许不同层面的个别价值观有所重叠。

| 层次 | 具体内容 | 内容 |
|---|---|---|
| 需求 | 马斯洛需求 | 生理需求、安全需求、爱和归属感需求、尊重需求、自我实现需求 |
| 动机 | 施瓦茨动机 | 安全、自我导向、刺激、传统、成就、遵从、权力、慈善、大同主义、享乐主义 |
| 共同价值 | 全人类共同价值观 | 和平、发展、公平、正义、民主、自由 |
| 特色价值 | 中国传统价值观 | 仁、义、礼、智、信、忠、孝、廉、耻、勇 |
| | 中国社会主义核心价值观 | 富强、民主、文明、和谐、自由、平等、公正、法治、爱国、敬业、诚信、友善 |
| 应用 | 价值正负向判断 | 正向、负向 |
| | 使用场景 | 国家、社会、个人、家庭 |

Table 1: 多维价值观体系与具体价值观

## 4 资源库构建

### 4.1 数据来源与筛选

谚语类工具书选择标准既定位为权威性和规范性，又兼顾实用性和多样性，力求体现收语数量的涵盖性和读者定位的普适性。受不同收录标准以及谚语本身界定的差异性，导致各类词典收录条数相差较大，但无论多少，都无法展现谚语的全貌。在考虑到词典的权威性和实用性等因素后，本文选取《新华谚语词典》、《中华谚语大词典》以及《Oxford Dictionary of Proverbs》作为谚语数据来源之一。

由于谚语词典中的副条目与主条目含义相似，本文主要选取并分析词典中的主条目。随后，删除重复的谚语和无明显价值观的谚语。经此处理后发现所筛选到的含明显价值观的西方谚语数量太少。因此，又利用美国当代英语语料库（Corpus of Contemporary American English，简称COCA [2] ）进行西方谚语的补充。

围，在多维价值体系中做到中西方价值观的平衡。

[2]是当今世界上最大的在线免费英语平衡语料库，网址：https://www.english-corpora.org/coca/


(c) 2024 中国中文信息学会计算语言学专业委员会

### 4.2 标注流程

在进行谚语价值观的标注时，GPT-4与人工标注流程在初始阶段基本一致，首先需要判定所收集谚语的需求和动机层次。其次，对于每个价值观内容，都将依据全人类共同价值观、传统价值观、社会主义核心价值观三个体系依次对谚语进行判断和分类。然后，标注工作需要明确谚语价值观的正负向及所使用的场景。在人工标注过程中，标注人员需要对标注结果进行最终检查。标注流程如图2所示。

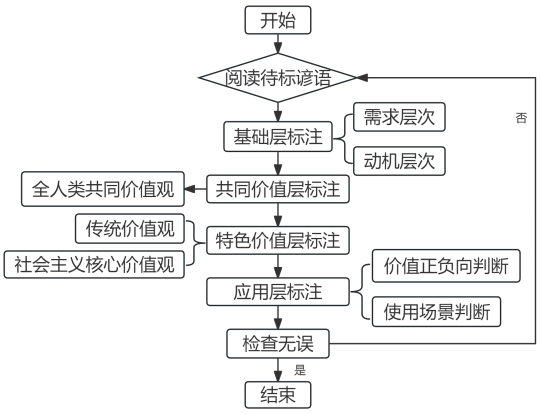



Figure 2: 标注流程

### 4.3 标注规范与一致性

为尽可能保证标注结果的正确性，不会影响后期的对比分析，在进行人工检验之前，我们制定了相应的标注规范。在标注人员的选择上，我们招募了三位语言学相关专业的在校研究生作为标注人员。在标注过程中，本文要求标注人员根据谚语在词典中的本义及其比喻义等来推断动机、需求层次及其所蕴含的价值观等。对于那些难以直接推断的谚语，将借助CCL语料库和BCC语料库，根据其在具体语境中的使用情况进行判断。

由于整个标注过程较为复杂，我们将标注分为三个环节，分别是培训环节、试标注环节以及正式标注环节。在培训环节，我们将介绍标注的理论背景、标注规则和标注流程。标注人员在熟悉标注规范之后，会进入试标注环节，在这一阶段中，标注人员将对100条语料进行试标注，试标注正确率低于80%的标注人员将被劝退。在正式标注环节，每条谚语将由两位标注员进行标注，对于不一致的谚语会交由第三人检查，期间出现的标注问题，也会由大家共同讨论决定，标注的总体一致率为82.2%。

### 4.4 标注结果

最终，数据集包含中西谚语2278条。其中中国谚语有1477条，占比65%，西方谚语有801条，占比35%。

| | 条数 | 占比 | 标注标签总数 | 每条谚语平均标签数 |
|---|---|---|---|---|
| 中国谚语 | 1477 | 65% | 7468 | 5 |
| 西方谚语 | 801 | 35% | 3815 | 4 |

Table 2: 标注结果统计

## 5 分析与讨论

### 5.1 基础层

#### 5.1.1 需求层

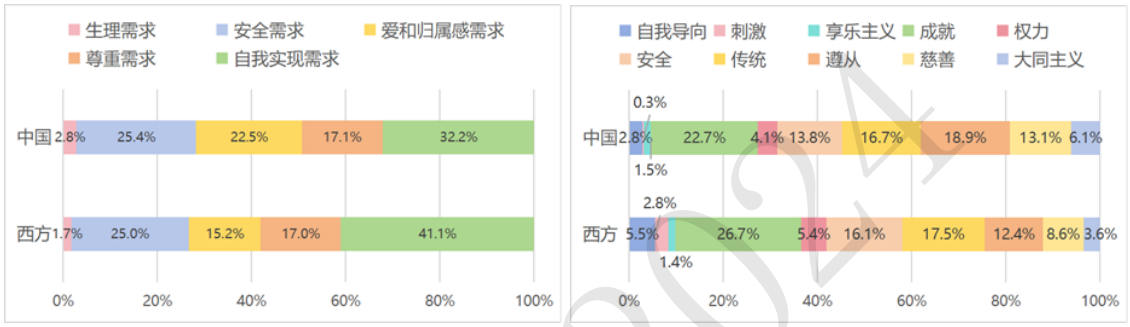
中西方谚语所体现的需求层次在各自文化中的占比如图3a所示。中国谚语在马斯洛需求层


(c) 2024 中国中文信息学会计算语言学专业委员会

次中占比最高的三个层次依次为“自我实现需求”、“安全需求”和“爱和归属感需求”。西方谚语则分别是“自我实现需求”、“安全需求”和“尊重需求”。

可以发现中西谚语都更多地关注马斯洛需求层次中的更高层次——自我实现需求，且西方重视程度更高，如“Where there’s a will, there’s a way”，强调了意志力在追求自我实现过程中的关键作用。“Actions speak louder than words”，强调了实际行动在实现自我价值中的核心地位。中国谚语中也有类似的表达，如“成大事者，不矜细行”强调在追求伟大事业或重要目标的过程中，人们不应过分拘泥于小节或琐碎的事务。“吃得苦中苦，方为人上人”鼓励人们不畏艰难困苦，勇于面对并克服挑战，以此来实现更高的目标或境界。而且中西方谚语在生理需求上的表达均较少，原因主要在于谚语所反映的价值观更多地关注精神层面和社会规范，而不是直接描述生理需求。但这并不意味着谚语与生理需求完全脱节，如“民以食为天”、“When the well is dry, we know the worth of water”等都表达了人民群众的生理需求。

除此之外，中国社会更加强调“爱和归属感需求”，谚语如“杯水之恩，江河还报”、“黄金难买乡邻情”。传统中国社会是一个“乡土社会”(费孝通, 2008)，是一个以血缘、地缘为基础的紧密社会网络。在这个网络中，人们往往会通过与亲人、朋友、邻居等建立紧密的联系，来获得情感上的支持和归属感，寻求在群体中的认同和接纳。西方谚语中也有类似的表达，如“What a neighbour gets is not lost”，但数量较少。




(a) 需求占比图　　(b) 动机占比图

Figure 3: 需求及动机层次理论标注结果占比图

### 5.1.2 动机层

中西方谚语所体现的动机层次在各自文化中的占比如图3b所示。可以发现中国谚语在施瓦茨动机层次中占比最高的三个层次依次为“成就”、“遵从”和“传统”。西方谚语则分别是“成就”、“传统”和“安全”层次。两者在整体分布上呈现出一定的相似性，中西方谚语所体现的“成就”层次均居于首位，谚语如“三百六十行，行行出状元”、“有志者事竟成”、“笨鸟先飞”、“Business before pleasure”、“Never say never”，说明在中西文化中都强调了成功、努力和决心等的重要性，体现了人类对于成就的共同追求。

中西谚语在“遵从”层次上的分布差异尤为明显，中国谚语在遵从层次上占比为18.9%，而西方谚语在遵从层次上占比为12.4%，二者占比相差6.5%。而且在层次内涵表达上，二者也存在差异。例如：在“自我导向”这一层次中，中西方谚语反映了各自文化对个人主义和自我实现的不同理解和重视程度。西方文化传统上更加强调个人主义和自我实现。谚语如“Be what you would seem to be”、“Every man for himself”都体现了这种文化倾向。这些谚语强调个人的独立性、自主性和对个人价值的追求，反映了西方文化中个人自由和自我实现的重要性。相较于西方，中国社会更加强调个体责任，谚语如“自己的梦自己圆”、“教的言语不会说，有钱难买自主张”，往往与集体利益和社会责任相结合。这些差异体现了中西方文化在价值观念和社会结构上的不同。

## 5.2 共同价值层

含全人类共同价值观的中国谚语共计366条，西方谚语共计148条。中西方谚语所体现的全人类共同价值观在各自文化中的占比如图4所示。中国谚语在全人类共同价值观体系中占比最高的价值观依次为“发展”、“正义”和“公平”，西方谚语占比最高的则为“发展”、“公平”和“正义”。


(c) 2024 中国中文信息学会计算语言学专业委员会

可以发现涉及“发展”的表达均居于首位，反映了人们对进步和改善现状的渴望。例如：“没有翻不过去的火焰山”、“开卷有益”、“Rome was not built in a day”、“We must learn to walk before we can run”。这些谚语鼓励个体和社会不断前行，强调的是通过个人和集体的努力实现社会和个人的发展和进步。其次，涉及“公平”、“正义”概念的表达也居于前列，例如，中文谚语中有“大丈夫见义勇为”、“为人坐得正，不怕影子斜”、“人不论大小，马不论高低”等，而在西方谚语中则有“Justice delayed is justice denied”、“One law for the rich and another for the poor”等表达。这一现象反映了人类社会普遍存在的对公平、公正的向往和追求。可以说谚语不仅是生活智慧的传递者，也是社会不公和不平等现象的批判者。

在“和平”这一维度上，中西方谚语虽然在占比上基本相同，但在具体内涵上却存在一些差异。在西方文化中，谚语关于和平的表达往往呈现出一种“实用主义”的色彩。西方谚语如“If you want peace, you must prepare for war”，强调的是一种防御性和战略性的和平观，即和平需要通过实力和备战来维护。这种观念反映了西方历史上国家间长期的冲突和战争经历，以及对国家安全和主权的重视。相比之下，中国文化在谚语中对和平的表达通常强调的是避免冲突，例如，“二虎相斗，必有一伤”、“有手不打上门客”，这些都体现了一种追求和平解决冲突、避免不必要损害的思想，并且这种和平观念与中国传统的集体主义、和谐与社会稳定的价值观紧密相关。总体来说，这些差异的形成深受各自文化背景和历史发展的影响。西方历史上经历了多次大规模战争，这些经历在一定程度上塑造了西方文化中关于和平的实用主义观念。而中国长期的儒家文化背景强调“和为贵”，这导致了中国谚语更加强调对和平的重视和推崇。

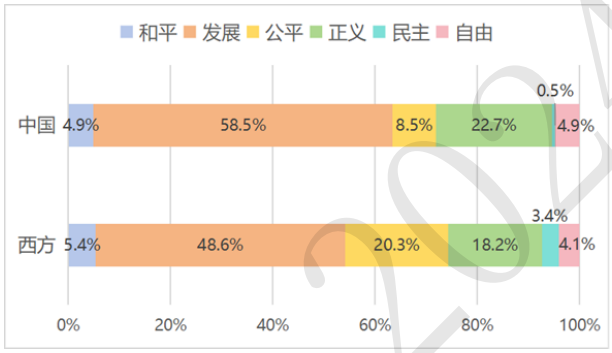



Figure 4: 全人类共同价值观标注对比图

## 5.3 特色价值层

### 5.3.1 传统价值观

含传统价值观的中国谚语共计832条，西方谚语共计414条。中西方谚语所体现的传统价值观在各自文化体系中的占比如图5所示。中国谚语在传统价值观体系中占比最高的价值观依次为“仁”、“礼”和“智”，西方谚语占比最高的则为“智”、“仁”和“义”。

中国谚语在“仁”维度上频率最高，“仁”主要是指“以人为本”。在中国文化中，尤其是受儒家思想的影响，仁的概念被赋予极高的价值。“仁”的内涵都是积极的，都是善良而美好的道德品质，其核心思想是“爱人”、“己所不欲，勿施于人”等(刘兆伟, 2015)。中国谚语在表达“仁”时，更强调道德积累和内在修养的重要性。如“积金不如积德，克众不如济人”，强调的是积累德行比积累财富更重要，帮助他人比胜过他人更有价值。“救人一命，胜造七级浮屠”也强调了救人性命、行善积德的崇高价值。而西方谚语在表达类似观念时，更注重个体与社会的互动关系以及行为的具体效果。例如“A lie that heals is better than the truth that hurts”，并不是在直接强调道德行为的崇高性，而是关注在特定情境下，善意的谎言可能比残酷的真相更有益。又如“Charity begins at home”更侧重于从家庭开始培养这种品质。但无论是中国还是西方，谚语在表达“仁”这一概念时都在提醒人们要关注他人的需要，积极行善。

西方谚语在“智”的维度上频率最高，且中西方相差较大。“智”指的是崇智求真，西方谚语如“All that glitters is not gold”、“Experience is the best teacher”等，强调了实证和经验的重要性。当然，中国谚语中也不乏有类似的表达，如“百闻不如一见”、“眼过千遍，不如手过一遍”等，这表明无论在东方还是西方文化中，谚语作为传播价值观的一种方式，反映了人们对真理和智慧的普遍渴望。


(c) 2024 中国中文信息学会计算语言学专业委员会

相反，“耻”的层面在中西方谚语中出现的频率相对较低。“耻”主要是指“人之为人的底线”，例如：“人怕揭短，龙怕揭鳞”。由于直接表达“耻”可能被视为不礼貌或过于尖锐，因此在中西方文化中表达此类概念时通常较为含蓄和间接。再者，由于谚语通常用于传达生活智慧、提供指导或反映普遍经验，而不是直接表达道德失范行为，这可能是“耻”在谚语中出现频率较低的原因。

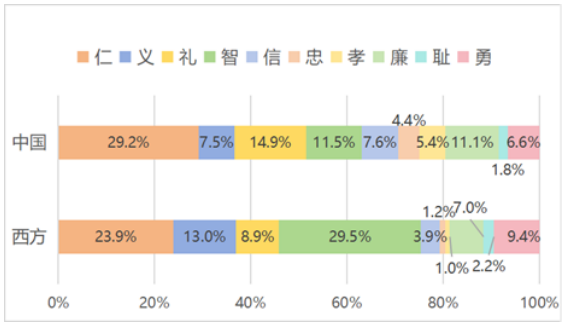



Figure 5: 传统价值观对比图

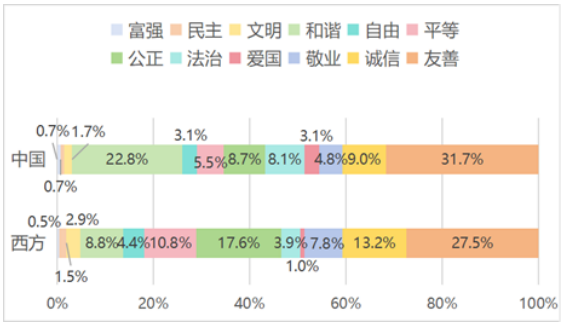



Figure 6: 社会主义核心价值观对比图

### 5.3.2 社会主义核心价值观

含社会主义核心价值观的中国谚语中共计543条，西方谚语中共计204条。中西方谚语所体现的社会主义核心价值观在各自文化中的占比如图6所示。中国谚语在社会主义核心价值观体系中，占比最高的价值观依次为“友善”、“和谐”和“诚信”，西方谚语占比最高的则为“友善”、“公正”和“诚信”。

在中西谚语中，“友善”这一概念出现频率均居于第一位，“敬人者，得人恒敬”、“贫穷患难，亲戚相救；婚姻死丧，邻里相助”、“和气生财”、“One hand washes the other”、“Least said, soonest mended”。它们都强调了人与人之间的相互帮助和尊重，这反映出中西方文化在基本价值观上的共通性，同时也反映了这些价值观在社会生活和文化传承中的重要性。

除此之外，中西谚语在社会主义核心价值观“和谐”这一维度上相差较大，中国谚语在“和谐”这一维度上占比为22.8%，西方谚语占比为8.8%。而且在具体表达上，二者也有一些不同，中国谚语更多强调家庭和社会的和谐与稳定，谚语如：“南北一家，兄弟一堂”、“家和万事兴”、“父子不和家不旺，邻居不和是非多”，西方谚语中除有类似的表达之外，还特别关注武力、权力等，以谋求社会的和谐。如：“Force, no matter how concealed, begets resistance”。在“诚信”这一具体表达上，二者也存在一些不同。例如西方谚语“An Englishman’s word is his bond”体现了诚实和信用的重要性，强调个人的承诺和信誉是不容置疑的。类似地，中文谚语“实话好说，谎话难编”、“话一出口，驷马难追”则强调言语的重要性。

但无论在哪种文化背景下，这些价值观都反映了人类社会的共同追求，也是社会稳定和文化传承的重要基石。

## 5.4 应用层

### 5.4.1 谚语价值判断

价值判断标注结果如图7所示，可以发现：中西方谚语价值观分布情况基本一致，正向价值观占比大于负向价值观。这主要是因为谚语作为传递社会价值的重要载体,其多数内容反映了社会和文化中推崇的正向价值观，从而能够在社会成员间发挥着积极的引导作用，这也证明了谚语富有教育意义(黄伯荣and 廖序东, 2017)。而且随着时间的推移和社会的发展，那些不符合时代精神的谚语往往会被淘汰或修正，而那些具有普遍指导意义的正向价值观则会得到更加广泛的传播和应用。谚语如“大伙心齐，泰山能移”、“A cat may look at a king”。而谚语中负向价值观的表达往往源于历史传统。例如，“金钱能使鬼推磨”这一中国谚语反映了历史上金钱至上、道德沦丧的负面观念，西方谚语中的“A woman’s place is in the home”则体现了历史上对女性角色的刻板印象和限制。

### 5.4.2 谚语价值使用场景

谚语是人民群众表现实际生活经验或感受的一种“现成话”(武占坤and 马国凡, 1980)，通常


(c) 2024 中国中文信息学会计算语言学专业委员会

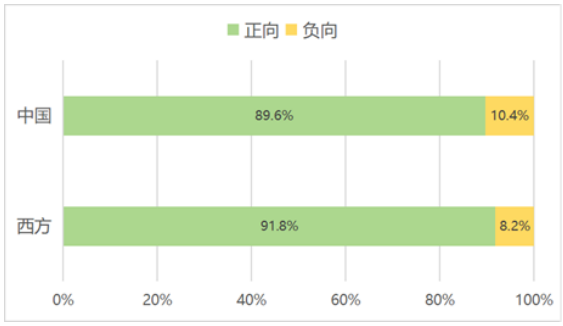



Figure 7: 价值判断对比图

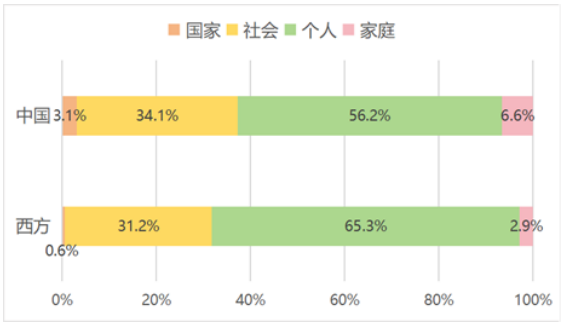



Figure 8: 使用场景对比图

源自个人的日常生活经验和智慧积累，反映了个体在特定社会和文化背景下的行为、态度、选择及其后果。因此，在中西方谚语中，个人层面的价值观所占比例最高，分别达到56.2%和65.3%，这反映了谚语在传递个人经验和智慧方面的核心作用，如图8所示。谚语如：“君子一言，驷马难追”、“A penny saved is a penny earned”等。其次，社会层面的价值观在谚语中也占据显著位置。谚语如：“公道自在人心”、“A friend in need is a friend indeed”。这表明谚语不仅是个人智慧的结晶，也是社会规范和道德准则的传播媒介。中西方谚语价值判断和使用场景的相似性也揭示了不同文化之间的共通性。

## 5.5 体现在需求和动机层上的谚语价值

### 5.5.1 体现在需求层上的谚语价值

各个价值在需求层次上的分布情况如图9所示。从中西方内部价值观分布来看，中国谚语在“自我实现需求”上占比最高，西方谚语在“尊重需求”上占比最高。因此，中国社会更多强调“自我实现需求”，而西方社会更多强调“尊重需求”。从同一个价值观中西方分布差异来看，在传统价值观中，中国谚语更多强调“爱和归属感需求”，西方谚语更多强调“尊重需求”。在社会主义核心价值观中，中国谚语更多强调“爱和归属感需求”，西方谚语更多强调“自我实现需求”。在全人类共同价值观中，中国谚语更多强调“自我实现需求”，西方谚语更多强调“尊重需求”。所以从这一角度来说，中国社会更多强调“爱和归属感需求”，而西方谚语更多强调“尊重需求”。这与中西方的历史传统有着紧密的联系，中国社会强调亲情、友情、爱情等多种形式的情感纽带。而在西方社会中，尊重被视为一种基本的人权和社会准则，这反映了西方文化对个体尊严和自我价值的高度重视。此外，各个价值观的来源更多的是代表“爱和归属感需求”、“尊重需求”和“自我实现需求”，也说明随着历史的发展，人民群众已经满足了最低需求，人们正逐步迈向对更高层次需求的追求。

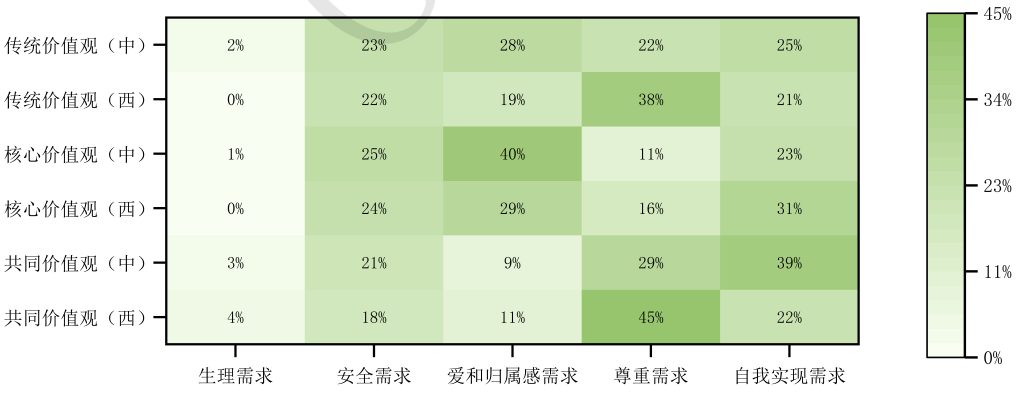



Figure 9: 价值观来源马斯洛需求层次体系分布图[3]

[3]颜色深浅的变化能够反映出价值观在某一动机或需求层次中所占的比重大小，颜色越深，则表明该价值观来源


(c) 2024 中国中文信息学会计算语言学专业委员会

### 5.5.2 体现在动机层上的谚语价值

各个价值在动机层次上的分布情况如图10所示。可以观察到来自传统价值观的中国谚语主要归属于“遵从”层次，西方谚语主要归属于“成就”层次。这主要是因为在中国传统文化中，更多强调对既有社会规范和道德准则的遵从。而西方谚语更多强调“崇智求真”，个人通过智慧和努力实现自我价值和成就。这与传统价值观的标注结果显示一致。其次，来自社会主义核心价值观的中国和西方谚语均主要归属于“遵从”和“慈善”层次。这与社会主义核心价值观提倡维护社会秩序以及为人“友善”有着密切的关系。来自全人类共同价值观的中国和西方谚语均主要归属于“成就”层次。这可能因为全人类价值观中的“发展”观与施瓦茨动机层次中的“成就”层次之间存在一定的内在联系。个人的“成就”动机可以在一定程度上推动社会的发展，同时，社会的发展也使个体能够更好地实现个人的成就和价值。整体来看，各个价值来源在施瓦茨动机层次体系分布中，“成就”、“遵从”、“慈善”层次居于前列，而“刺激”、“享乐主义”居于末端。

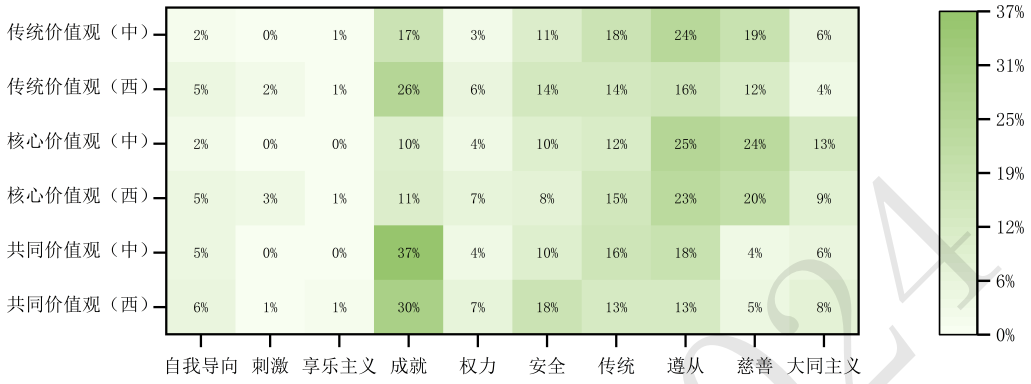



Figure 10: 价值观来源施瓦茨动机层次体系分布图

### 5.6 价值取向

价值取向是指人们把某种价值作为行动的准则和追求的目标。它是个体的活动或意识中所渗透的价值取向，是人们实际生活中追求价值的方向。词汇的使用在一定程度上能够反映谚语价值观的表达，经去除停用词、分词等数据预处理工作后，统计了词频，中西方谚语词云如图11、12所示。其中，中西方谚语出现频率最高的词分别是“人”、“man”。说明不管是中国还是西方，都强调以人为本的价值取向。中国谚语中共有421条谚语中包含“人”这一核心要素，如“帮人一口得一升，救人一命积善功”、“不听好人言，必有惶泪”、“撑船撑到岸，助人助到底”，西方谚语中共有71条反映了“人”这一要素，如“A wise man will learn from his mistakes, but a fool will learn no other way”。并且大部分覆盖了本文所标注的价值类别，这表明虽然文化背景和价值观存在差异，但中国和西方都非常重视人性、人际关系、人的价值，寻找人类共同的价值追求。

在对中西谚语共现词进行统计后，发现中国谚语尤为强调“家”与“父母”的重要性。在中国谚语中，“家”和“父母”出现频率为88，谚语如：“不当家不知柴米贵”、“当家才知柴米价，养儿方晓父母恩”、“父母恩比天大”，深刻揭示了家庭责任与孝道的内涵。相比之下，在西方谚语中出现的频率仅为12。谚语如“The family that prays together stays together”。此外，西方谚语更多关注“神”的角色与意义。西方谚语中“god”出现频率为20，谚语如：“God helps them that help themselves”、“In the child’s mouth and heart, the mother is god”、“Trust in god but tie your camel”、而在中国谚语中出现的频率仅为10，谚语如“举头三尺有神明”等。这些谚语反映了中西文化的差异。

## 6 总结

人民创作是研究人民心理的非常必需而重要的材料。谚语作为一种传统的智慧和知识的传

这一动机或需求层次更多。


(c) 2024 中国中文信息学会计算语言学专业委员会




Figure 11: 中国谚语词云图

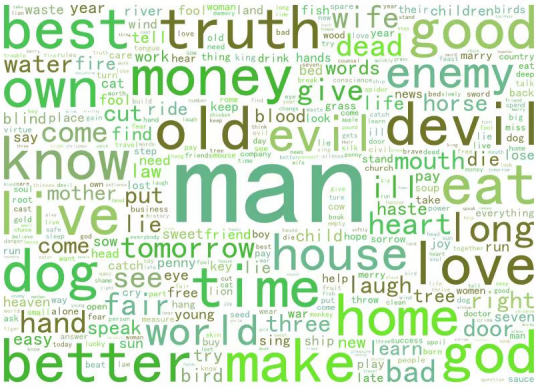



Figure 12: 西方谚语词云图

播方式，反映了文化中所推崇的价值观。本文在构建多元价值层次谚语资源库的基础上，对谚语进行了多角度的对比分析。我们发现：

中西方谚语所蕴含的价值观及在表达上展现出共通之处。中西谚语中所反映的正向价值观的数量大于负向价值观的数量，且谚语中体现的价值观在个人和社会层面上占据显著比重。在多元价值观表达上，二者也存在相似性，比如“发展”、“公平”、“正义”、“友善”等价值观出现的频率均较高。此外，中西谚语在“成就”动机层次中表达最为显著，在需求层次上都注重追求马斯洛需求层次中的更高层次。在价值取向表达上，中西双方都关注“人”的价值。

中西方谚语也彰显出各自独特的文化特色。西方谚语在“智”的维度上频率最高，而中国谚语在“仁”这一维度上频率最高，且中国谚语在表达“仁”时，更强调道德积累和内在修养的重要性，而西方谚语更注重个体与社会的互动关系以及行为的具体效果。在“和平”这一价值观表达中，中国文化凸显出解决冲突、避免不必要损害的思想，西方文化更多强调和平需要通过实力和备战来维护。在“自我导向”层次中，中西文化展现了各自文化对个人主义和自我实现的不同理解和重视程度。

通过比较分析中西谚语中的价值观，我们能更好地理解不同文化之间的共通性和差异性，促进不同文化之间的相互理解和尊重。本文从动机层、需求层、共同价值层、特色价值层、应用层多个层次对中西方谚语进行了描述和分析，考察了各个价值与施瓦茨动机层次及马斯洛需求层次之间的关系，并研究了谚语词汇的使用特点，与价值表达的关系，是对中西价值观对比分析的一次新的尝试。但是，本文也存在一些不足之处有待进一步改进。首先，本文虽然采用了GPT-4和人工辅助的双重标注方法，以增强数据的准确性，但标注结果仍可能受到一定程度的主观性影响，需要进一步校正。其次，受谚语词典收录标准的不同，词典编纂者在选择收录的谚语时，可能更倾向于那些与人类社会活动和个人行为密切相关的内容，这种选择偏好可能导致了某一类别收录数量或多或少。再者，目前研究覆盖的中西方谚语数量有限，未能全面反映两种文化中谚语的价值取向。未来研究将继续扩大谚语资源库的规模，以便更加全面深入地揭示中西方谚语中的价值观。


(c) 2024 中国中文信息学会计算语言学专业委员会

## 参考文献

Ryan Boyd, Steven Wilson, James Pennebaker, Michal Kosinski, David Stillwell, and Rada Mihalcea. 2015. Values in words: Using language to evaluate and understand personal values. *Proceedings of the Ninth International AAAI Conference on Web and Social Media*, 05.

Jilin Chen, Gary Hsieh, Jalal U. Mahmud, and Jeffrey Nichols. 2014. Understanding individuals' personal values from social media word use. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '14, page 405–414, New York, NY, USA. Association for Computing Machinery.

Emi Ishita, Douglas W. Oard, Kenneth R. Fleischmann, An-Shou Cheng, and Thomas Clay Templeton. 2010. Investigating multi-label classification for human values. In *Proceedings of the 73rd ASIST Annual Meeting on Navigating Streams in an Information Ecosystem - Volume 47*, ASIST '10, USA. American Society for Information Science.

Pengyuan Liu, Sanle Zhang, Dong Yu, and Lin Bo. 2022. CoreValue:面向价值观计算的中文核心价值-行为体系及知识库(CoreValue: Chinese core value-behavior frame and knowledge base for value computing). In Maosong Sun, Yang Liu, Wanxiang Che, Yang Feng, Xipeng Qiu, Gaoqi Rao, and Yubo Chen, editors, *Proceedings of the 21st Chinese National Conference on Computational Linguistics*, pages 417–430, Nanchang, China, October. Chinese Information Processing Society of China.

Abraham H. Maslow. 1943. A theory of human motivation.

Liang Qiu, Yizhou Zhao, Jinchao Li, Pan Lu, Baolin Peng, Jianfeng Gao, and Song-Chun Zhu. 2022. Valuenet: A new dataset for human value driven dialogue system. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11183–11191, June.

Shalom Schwartz and Wolfgang Bilsky. 1987. Toward a universal psychological structure of human values. *Journal of Personality and Social Psychology*, 53:550–562, 09.

W.D.拉蒙特. 1992. 价值判断. 北京：中国人民大学出版社.

侯洁诚. 2017. 英汉谚语对比分析. 语文建设, (29):69–70.

刘兆伟. 2015. 论语. 北京：人民教育出版社.

戴木才. 2015. 中国人的美德与核心价值观. 北京：中国人民大学出版社.

朱益平. 2009. 从汉英谚语对比看中西价值观差异. 理论导刊, (12):120–121+119.

武占坤and 马国凡. 1980. 谚语. 呼和浩特：内蒙古人民出版社.

王德春. 2003. 汉英谚语与文化. 上海：上海外语教育出版社.

费孝通. 2008. 乡土中国. 北京：人民出版社.

郭绍虞. 1925. 谚语的研究. 上海：商务印书馆.

钟鸣and 代玮炜. 2008. 中西方谚语对比研究中折射出的中西方文化异同. 西南民族大学学报*(*人文社科版*)*, 29(S1):112–114.

马利军and 张积家. 2017. 汉语谚语语义性质及其互相作用关系研究. 浙江大学学报*(*人文社会科学版*)*, 47(02):42–54.

黄伯荣and 廖序东. 2017. 现代汉语. 北京：高等教育出版社.

黄希庭, 张进辅, and 李红. 1994. 当代中国青年价值观与教育. 成都：四川教育出版社.

CCL 2024

# 从句子图到篇章图——基于抽象语义表示的篇章级共指标注体系设计

**张艺璇** [1,2], **李斌** [1,2,3], **许智星** [1,2], **卢芃秀** [1,2]
1 南京师范大学文学院, 江苏南京 210097
2 南京师范大学语言大数据与计算人文研究中心, 江苏南京 210097
3 澳门大学人文学院, 中国澳门 999078
xiaoxian224@gmail.com, libin.njnu@gmail.com,
xzx0828@live.com, 230102030@njnu.edu.cn

## 摘要

篇章共指体现篇章概念的动态转移，成为近年研究热点。本文在梳理共指理论研究的基础上，综述了相关语料库及解析方法，发现共指语料库仍存在以下两个问题：共指关系标注粗疏与基本不考虑整句语义表示的融合。本文以句子级语义标注体系（中文抽象语义表示）为基础构建篇章共指体系，构建了 100 篇共指语料库。本体系涵盖 52 种句内语义关系和 8 种篇章共指关系，二者相结合构建的篇章共指语义图，为篇章级语义分析提供新的框架和数据资源。

**关键词：** 篇章共指 ；抽象语义表示 ；篇章语义结构 ；中文信息处理

# From Sentence Graphs to Discourse Graphs: Designing a Discourse-Level Anaphora Annotation System Based on Abstract Semantic Representation

**Yixuan Zhang**[1,2], **Bin Li**[1,23], **Zhixing Xu**[1,2], **Pengxiu Lu**[1,2]
1 School of Chinese Language and Literature, Nanjing Normal University, China
2 Center for Language Big Data and Computational Humanities, Nanjing Normal University,China
3 Faculty of Arts and Humanities, University of Macau, China
xiaoxian224@gmail.com, libin.njnu@gmail.com,
xzx0828@live.com, 230102030@njnu.edu.cn

## Abstract

Discourse coreference reflects the dynamic conceptual transitions within a text and has become a research focus in recent years. This paper, based on a review of coreference theory, summarizes related corpora and parsing methods, identifying two main issues in existing coreference corpora: coarse annotation of coreference relationships and a lack of integration with complete sentence semantics. Using a sentence-level semantic annotation system (Chinese Abstract Meaning Representation) as a foundation, this paper constructs a discourse coreference system and a corpus of 100 texts. This system encompasses 52 types of intra-sentence semantic relationships and 8 types of discourse coreference relationships. The combination of these elements constructs a discourse coreference semantic graph, providing a new framework and data resource for discourse-level semantic analysis.

**Keywords:** Discourse Anaphora , Abstract Semantic Representation , Discourse Semantic Structure , Chinese Information Processing

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

## 1 引言

指代（Reference）不仅表现为语法上的替代关系或语义上的同指关系，还充当着话题转换的衔接手段，对挖掘篇章概念转移、理解自然语言有重要意义，目前已成为理论语言学和计算语言学关注的热点问题之一。

鉴于研究目的与分析方法的差异，目前理论语言学和计算语言学在指代领域的术语使用有所区别。理论语言学常使用“回指（Anaphora）”来表示指代现象，广义的回指可分为直接回指和间接回指两类 (Ariel, 1990)。直接回指是指两个语言成分之间的关系，对其中一个成分的解释，取决于对另一个成分的解释 (Huang, 1984)，间接回指是指回指语和先行语之间没有明显的指代关系，而需要经过语用推理才能建立指称关系 (Erkü and Gundel, 1987)。在计算语言学中，与“直接回指”相对的是“共指（Coreference）”，与“间接回指”相对应的是“桥接关系（Bridging Relation）”。但间接回指（或桥接关系）由于其复杂性与模糊性，被学界研究长期排除在外，故相关研究发展缓慢且未成体系。尽管理论语言学界已有学者 (王军, 2004; 王军, 2013; Vieira and Poesio, 2000) 尝试总结常见间接共指类别，计算语言学界也构建了一系列桥接关系语料库，如 ISNotes (Markert et al., 2012)、BASHI (Rösiger, 2018)、ARRAU (Poesio et al., 2008; Uryupina et al., 2016) 和 SciCorp (Rösiger, 2016)。但并没有对该现象的本质上达成一致认同，故目前本文仅讨论具有概念同一性的共指现象。

为推动共指消解（Coreference Resolution）任务的发展，计算语言学界开始标注共指语料库。尽管现有语料库的构建体系和标注方法已有一定发展，但仍存在以下两个问题，即共指关系标注粗疏以及基本不考虑整句语义表示的融合，导致共指消解仍有较大发展空间。因此，本体系将在中文抽象语义表示（Chinese Abstract Meaning Representation；CAMR）基础之上构建篇章共指标注体系。CAMR 是句子级标注体系，有 5 种核心语义角色关系和 47 种非核心语义角色关系，可有效挖掘共指词所担任的语义角色，提供语义关系。本体系以“概念同一性”为基本原则，从词形、词长和词类等角度区分 8 种共指关系，提供共指链内部、共指词之间的共指关系。先行词、共指词以及句内语义关系与篇章共指关系相结合可构建出篇章共指语义图，从语义角度归纳共指规律。基于本体系所构建的 100 篇共指语料库，可为共指解析提供数据支持。

## 2 直接回指的理论研究

已有大量学者从句法、认知、功能、篇章等角度对回指理论作出探索，其中影响深远的是认知角度的可及性理论（Accessibility Theory) (Ariel, 1990) 以及篇章角度的话题延续性模式（Topic Continuity Model) (Givón, 1983)。在两大理论提出的“可及性标示阶”与“话题延续性量表”的基础之上，学界逐步确定了频次、位置、间隔距离、语篇结构、干扰数量和句法成分等回指计量指标，但这些指标大多从表层分析回指，较少有学者构建回指的语义语料库，从语义角度归纳回指规律。纵观回指计量发展历程，研究焦点主要有三大趋势：其一，逐渐从名词和代词等具体实体的标注与计量 (徐赳赳, 2003; 许余龙, 2004; 宋宏, 2010; 石艳华, 2014) 转向概述回指等抽象类别 (刘东虹, 2014; 王秀丽, 2012; 吴丹苹, 2021; 寇鑫 and 徐坤宇, 2023)；其二，近年来，学界已不再局限于单纯地专研某一回指类型，而是将其放入句法、句式等宏观视野中探索规律。如蒋平 (2021) 基于认知语言学总结存现宾语的零形回指演进，冉晨 (2024) 分析数量名回指语的指称性质与回指确认方式；其三，回指研究视角逐渐从本体研究扩展到语言教学与应用中。如杨永生 (2020) 基于回指确认考察韩国学生汉语“这/那”句的习得情况，李榕等 (2021)、曹娜等 (2024) 借鉴徐赳赳 (2003) 等学者提出的计量指标挖掘韩国留学生的回指偏误、归纳二语学习者使用主语回指语的显隐现象。

## 3 共指的计算资源构建与解析研究

计算语言学借鉴现有理论标注了名词、代词、零形式等常见体词共指，并进一步探索动词、形容词甚至副词等谓词共指，一定程度上推动了共指研究。但谓词共指存在意义实虚难以界定等问题，仍处于探索阶段，体词共指的标注仍为学界主流。但这些语料库的标注信息相对粗疏，不能考察共指规律，也不能为现有计算模型提供充足的共指语义或知识信息。

### 3.1 共指相关语料库构建

面向计算建模的共指语料库可分为两类：一类是以 OntoNotes 为代表的共指语料库，一类是以 MS-AMR 为代表的包含共指信息的语料库。前者细致标注共指现象，却忽略了整句的语义


(c) 2024 中国中文信息学会计算语言学专业委员会

表示；后者全面表示整句语义表示，对共指现象的标注却仍待深化。但这两类语料库共有的不足在于均仅停留在了共指的同一性标注中，没有对共指链中共指词与先行词之间的关系作进一步分析。

### 3.1.1 共指语料库

最为典型的共指语料库为 OntoNotes (Weischedel et al., 2013) 语料库，该语料库也是目前共指评测认可度最高的数据集。其标注规范成为众多共指标注体系的借鉴样本，使共指语料库迈向多语言时代。尽管语料库数量在不断增加，规模不断扩大，语种不断丰富，但标注信息和标注方法仍并未发生明显更新。除了该系列语料库，共指语料库还包括 MUC (Hirschman and Chinchor, 1998)、ACE (Doddington et al., 2004) 等早期语料库，另外还有一些标注单一共指信息的语料库，如汉语零指代语料库 (孔芳 et al., 2021)、法汉指称链条平行语料库 (胡霄钦 and 王秀丽, 2021)。这些语料库将共指标注从体词扩展到谓词，为探索篇章共指提供新的渠道。但这些语料库几乎不涉及共指的语义信息，也并未挖掘共指链内部的共指关系。MUC6（Message Understanding Conference）开创了共指标注体系的先河。尽管其标签分类较为简单，但研究学者已逐渐认识到共指的复杂性，并有了灵活表示共指词的倾向性（如设置“MIN”标签以表示包含在整词之内的部分字符串）。但其细则也存在值得商榷之处，如将人物与其头衔等职位标注为共指关系、不补充标注零形式等问题。ACE（Automatic Content Extraction）在 MUC6 的基础之上，对共指进一步标注。其语料不再局限于英语，而是迈向了多语言资源的构建，ACE 也是最早针对中文指代消解的国际评测语料资源。MUC6 和 ACE 对共指的标注较为简单，直到 OntoNotes 之后共指标注迎来飞跃。OntoNotes 取消对共指标注的限制，除名词、代词之外，还可标注动词、形容词等信息，对指代消解任务有新的推进。在共指层主要标注两类信息，即同一性（IDENT）和同位语（APPOS）。同一性共指（IDENT）用于指代共指，意味着代词、名词性和特定所指对象的命名提及之间的联系，不包括一般的、未指定的或抽象实体。同位语（APPOS）逻辑上代表属性，因此被单独处理。自 OntoNotes 问世之后，主要有两大研究现状与趋势：其一，在 OntoNotes 标注规范的基础之上构建了许多大型语料库。Ghaddar 等 (2016) 继承 OntoNotes 标注方法，以维基百科为语料构建了 WikiCoref 语料库；Chen 等 (2018) 在 OntoNotes 规范的基础之上改良并构建了目前最大规模的 PreCo 英语共指语料库；Poesio 等 (2019) 基于游戏化方法构建了众包共指语料库，是最大的应用共指语料库之一。其二，近三年的共指资源主要集中于非英语语言以实现跨语言共指解析。这些语料库包括俄语共指语料库 RuCoCo (Dobrovolskii et al., 2022)、涵盖英法德葡四种语言的多语言共指语料库 ParCorFull2.0 (Lapshinova-Koltunski and Ferreira, 2022)、基于 UD（Universal Dependencies）构建的多语言语料库集合和共指消解的标准化格式 CorefUD (Nedoluzhko et al., 2022) 以及荷兰语跨文档事件共指解析大型数据集 (De Langhe et al., 2023) 等数据资源。

### 3.1.2 融合共指信息的整句语义标注语料库

以 PDT (Mikulová et al., 2006)、UMR (Uryupina et al., 2016) 和 MS-AMR (O’ Gorman et al., 2018) 为代表的语料库实际是句子级语义标注语料库，并非专为共指而设计。尽管一定程度弥补了共指语料库缺乏整句语义信息的不足，但在共指规范上仍待完善。

PDT 的共指信息有三种，即语法共指、文本共指和特殊类型共指，主要标注了代词、动词、省略、概述共指等共指信息。UMR 是建立在 AMR 基础之上、具有语言包容性的语料库，除了继承了 AMR 的句子级表示之外，在共指方面采用关系三元组的形式标注实体共指和事件共指。MS-AMR 参考了 OntoNotes 的标注方式在句子级 AMR 基础之上扩展句间共指信息。

## 3.2 共指关系的计算

从数学角度而言，共指关系实质是一种等价关系，因此消解过程就是等价类划分的过程 (宋洋 and 王厚峰, 2015)。在过去的十年中，共指解析已经从基于规则的方法发展到基于学习的方法 (周炫余 et al., 2014)。基于规则特征的泛化能力较差，理解和实现比较简单 (Lang et al., 2007)，基于学习的方法开始引入开放知识作为额外特征，改善了模型效果高度依赖特征工程的弊端。近年来采用深度学习算法，如多层感知器/循环神经网络方法、基于知识的方法和基于变压器的方法 (Liu et al., 2023)，逐渐突破大规模语料的限制，大大增强了模型的深层语义学习能力和泛化性能 (陈远哲 et al., 2019)。Liu 等 (2023) 总结出五点共指解析的挑战及发展方向，即：下游任务缺乏带有共指解析标签的数据集；缺乏符号特征与子符号特征的组合；需结合现有语


(c) 2024 中国中文信息学会计算语言学专业委员会

言研究及认知直觉；当前模型需压缩所用资源以实现多任务学习；仍需超大规模语言模型的出现。Liu 指出，尽管已有大量研究成果为数据注释和模型设计奠定理论基础，但这些语言和认知发现却很少被纳入基于深度学习的共指解析模型中，未来可结合额外的符号特征（如语义特征和知识表示）和子符号方法（如词嵌入） (Mao et al., 2018; Cambria et al., 2022; Mao et al., 2022)，以解决共指数据稀疏问题，提高共指消解的精度。

## 4 篇章级共指标注体系

综上所述，理论语言学界用于计量分析的语料库，大多只能从表层信息（如回指分布、句法成分等）计量并分析回指规律，缺乏以语义为研究导向的语料库；计算语言学界用于共指消解的语料库，也缺乏兼顾语义关系与共指关系的语料库，以提高共指消解的精度。因此，本文基于句子级语义标注体系（中文抽象语义表示）构建篇章共指标注体系，涵盖语义信息与共指信息，为共指消解提供新思路。

### 4.1 句子级 CAMR 的标注体系

Abstract Meaning Representation（AMR）是一种抽象语义标注框架，采用单根有向无环图的表示方法，图中节点表示概念，边表示概念之间的关系 (Banarescu et al., 2013)。AMR 忽略语义较虚的成分（如冠词、单复数、时态等等）和形态变化，从原句中抽象出概念，不拘泥于原句词语，允许对其进行增添、删减和改动等操作以便更好表示语义关系。李斌等 (Li et al., 2019) 在 AMR 的基础之上结合汉语的语言特点继承并发展出 CAMR（Chinese AMR）。其中，共包括 5 种核心语义角色关系（arg0-arg4）、47 种非核心语义角色关系。

句子级 CAMR 体系是篇章共指标注体系的构建基础，以词为基础单位，以树结构为展示形式，语义结构的层级性有利于在复杂的语言表述中确定共指单位，并提供共指单位的语义功能。表 1 为句子级 CAMR 标注体系和篇章级共指标注体系比较。

| **体系名称** | 句子级 CAMR 标注体系 | 篇章级共指标注体系 |
|---|---|---|
| **数据形式** | 单根有向无环图 | 单根有向无环图 |
| **标注层级** | 句内（句子级） | 句内、句间（篇章级） |
| **语义单位** | 单句内的概念（汉语常以词表示） | 篇章内的共指概念（汉语常以词或短语表示） |
| **揭示信息** | 句内各个成分的语义结构 | 篇章内部共指成分的指代关系 |
| **文本示例** | 男孩 [1] 希望 [2] 女孩 [3] 相信 [4] 他 [5] 。[6]<br>x5/希望-01<br>:arg0() x1/男孩<br>:arg1() x5/相信<br>:arg0() x3/女孩<br>:arg1() x5/x1 | p0222_r3_entity<br>:root() s3_x55 / university（杭州大学）<br>:alias() s3_x26_x27 / 该校<br>:homo() s9_x24 / university（杭州大学）<br>:reduce() s13_x4 / 大学<br>:reduce() s13_x19 / 大学 |

表 1: 句子级 CAMR 标注体系与篇章级共指标注体系比较

备注：在文本示例中，句子级 CAMR 标注体系展示的是单句的语义结构，而篇章级共指标注体系的示例则展示了在同一个篇章中指代“杭州大学”的所有语言单位。左侧，形如“:arg0”的标签为语义关系，“x+ 数字”表示词编号；右侧，形如“:alias”的标签为共指标签，“s+ 数字”表示句编号，“x+ 数字”表示词编号。

### 4.2 共指标注原则

#### 4.2.1 概念同一性

本体系遵守“概念同一性”原则，只标注相同指代意义的词语。如表 2 文本中与“福利院”有关的词语共出现 7 次，但有 6 次（位于 S3、S6、S7、S8、S11）指代为“杭州市儿童福利院（位于 S3）”，构成共指链 2；有一次（位于 S4），泛指国内的福利院，在本文中与“中国孤儿院（S1）”形成共指关系，构成共指链 1。

备注：表 2 为 CTB8.0 语料中编号为 0222 的文本，为便于理解，省略了与“福利院”“孤儿院”无关的句子。


(c) 2024 中国中文信息学会计算语言学专业委员会

| |
|---|
| S1：丁豪成为 [**中国孤儿院** ]$_1$ 长大的第一个残疾人大学生<br>S3：自幼成为孤儿并在 [**杭州市儿童福利院** ]$_2$ 长大的残疾学生丁豪，日前收到了杭州大学的录取通知书，成为该校哲学与社会科学系的学生。<br>S4：据了解，今年十九岁的丁豪是新中国第一个由 [**福利院** ]$_1$ 收养成长的残疾孤儿大学生。<br>S6：三岁时一场高烧，使他患上了严重的小儿麻痹后遗症，这一年他被 [**福利院** ]$_2$ 收养。<br>S7：丁豪在 [**儿童福利院** ]$_2$ 读完小学，随后进入附近乡里一所学校上初中。<br>S8：一开始由 [**福利院** ]$_2$ 的老师接送，初二后住校，每个星期六回到 [**福利院** ]$_2$ ，无论刮风下雨，从不缺课。<br>S11：[**儿童福利院** ]$_2$ 十六年如一日地关心、爱护他，去年以来，仅为他交纳学费、买资料、买营养品，就花了三千多元人民币。 |

表 2: “杭州市儿童福利院”共指示例

#### 4.2.2 共指链可以嵌套

由于语言递归性，词语嵌套是语言表达的常用方式，因此极易出现共指词或共指链嵌套的情况。表 3 的例子共出现两条共指链，下标相同即为同一条共指链。S17 中的“其名”即存在共指词的嵌套现象：“其”指代“小海龟”，为共指链 2 的共指词；“其名”指代“卡洛塔”，为共指链 1 的共指词。当多个共指链的每一个共指词均存在嵌套现象时，则这些共指链存在嵌套现象。

| |
|---|
| S16：本届世界游泳锦标赛的吉祥物是一只名叫“ [**卡洛塔** ]$_1$ ”的 [**小海龟** ]$_2$ 。<br>S17： [**海龟** ]$_2$ 的形象由意大利知名图案设计师瓦·隆巴多设计，[[**其** ]$_2$ **名** ]$_1$ 取自于一位天真可爱的意大利小女孩。<br>S18：别看 [**小海龟** ]$_2$ 在沙滩上爬行往往显得笨拙，一旦到了海里，就变得灵活自如。<br>S19：在“[**小海龟** ]$_2$ ”坚实的龟甲上嵌着由陆地、海洋组成的世界地图，寓意来自全球五大洲四大洋的游泳健儿会聚罗马，通过比赛，加强了解，增进友谊。 |

表 3: “卡洛塔”小海龟例句

备注：表 3 为 CTB8.0 语料中编号为 0308 的文本，为便于理解，省略了与“海龟”“卡洛塔”无关的句子。每行的“S+ 数字”为句子编号。

#### 4.2.3 区分概念同一性和实体间关系

篇章内部存在大量指代不同、但关系密切的概念，极易混淆。但本体系严格区分这类现象，未来将总结该类现象设计普适性的实体间关系。如表 3 中存在的两条共指链，一条是实体链（“小海龟”链），一条是实体名称链（“卡洛塔”链），即为指代不同但关系密切的共指链。

### 4.3 共指关系类型

本体系标签综合词长、词形、词义和词类等角度，依据先行词和共指词的中心词和定语变化等指标设置标签类别，共 8 大类（见表 4）。由于“:encap”（概述共指）和“: illustrate”（分述共指）标签存在方向性，于表 5 中的“标签小类”分别作详细介绍。“示例”部分为当先行词为“两个美丽的女孩”且共指元素均指代这一女孩时，共指标签的适用情况。

## 5 共指语料库标注及计量分析

### 5.1 语料选择

我们从 CAMR v2.0 语料中选取了 100 篇文本进行共指标注，该文本在宾州中文树库（Chinese treebank；CTB）的编号为 chtb0222-chtb0323（chtb0319 因篇幅过短不予标注）。该语料库体裁涵盖经济、体育及生活，经数据清洗（去除记者、发表时间等新闻组成要素）后共有句子 1116 个，词次为 28303。由于该语料库同时拥有 CTB 句法信息和 CAMR 语义信息，可为共指标注与后期的计量检索提供极大帮助与便利。原因如下：其一，句子语义层级性以树结


(c) 2024 中国中文信息学会计算语言学专业委员会

| 编号 | 标签 | 含义 | 示例 |
|---|---|---|---|
| 1 | :root | 该词为先行词 | 两个美丽的女孩 |
| 2 | :homo | 共指词与先行词完全相同 | 两个美丽的女孩 |
| 3 | :add | 共指词在先行词基础之上添加某些词语 | 两个美丽善良的女孩 |
| 4 | :reduce | 共指词在先行词基础之上减少某些词语 | 两个女孩 |
| 5 | :alias | 共指词在先行词基础之上同时减少、添加某些词语（即改换其他表述） | 两个贤惠的女孩 |
| 6 | :pron | 共指词是代词 | 她们 |
| 7 | :illustrate | 分述共指 | —— |
| 8 | :encap | 概述共指 | —— |

表 4: 共指元素标签类型及含义

| 标签 | 标签小类 | 含义 |
|---|---|---|
| :illustrate | :illustrate | 共指词对先行词分述 |
| | :illustrate_of_alias | 共指词对先行词总述，共指词是代词 |
| | :illustrate_of_pron | 共指词对先行词总述，共指词是短语 |
| :encap | :encap_pron | 先行词是小句，共指词是代词 |
| | :encap_alias | 先行词是小句，共指词是短语 |
| | :encap_of_pron | 先行词是代词，共指词是小句 |
| | :encap_of_alias | 先行词是短语，共指词是小句 |

表 5: 共指部分标签小类及含义

构为展示形式，为共指词的确定提供精确定位；其二，CTB 涵盖分词、词性标注、句法等多种语言信息，涵盖学界现有的计量指标（如间隔距离、干扰数量、句法信息）；其三，CAMR 在 CTB 基础之上提供了语义信息，可揭示共指词变化的语义规律。

### 5.2 规范制定与标注过程

标注工作分为 3 个阶段。

（1）观察语料阶段。该阶段的主要工作是分析大量生语料，观察并标注出篇章中具有概念同一性的语言单位，在充分考虑语料质量与标注信息可操作性的基础之上，形成初步的标注规范。

（2）预标注阶段。该阶段将通过实践确认标注者对规范的理解，同时检验规范的通用性与可操作性，并在标注过程中及时调整规范，得到最终的标注规范。

（3）正式标注阶段。经过前期的标注工作，正式标注阶段采用了“程序粗提取 + 人工细标注”的方式，以词频为指标，在 CAMRv2.0 语料中提取潜在共指元素，建构粗糙的依存三元组，并在此基础调整修改，确保共指链标注无遗漏。最终形成完整的抽象语义表示篇章共指语料库。

### 5.3 语料计量分析

在这 100 篇文本中，共标注了 792 条实体共指链。每篇平均 7.92 条共指链，共指链长度处于 2-34 之间，平均每条共指链长度为 4.35，方差为 11.87，由此数据可以看出共指链长度变化极大，其中有 52% 的共指链长度处于 2-5 之间，其余共指链分布长度不等且分布不均；共指链跨句情况处于 0-34 之间，平均每条共指链跨 6.01 句，方差为 31.58；句内共指出现 509 次，句间共指出现 2937 次，平均每条链出现 0.64 次句内共指、3.71 次句间共指，方差分别为 1.73、7.11。经统计，各共指的出现频次见表 6（已按频次进行降序排列）。可以看到：首先，在共指关系中，共指词与先行词一致的情况占比最高，共出现 1495 次，但其方差也最高（6.88），说明这种共指关系出现次数最高，但在共指链中变化波动最大；其次，“:add”关系略低于“:reduce”关系，这可能与新闻语料的语言风格相关，即在论述一个概念时，新闻通常在首次提及时便提供最为详细的信息；最后，对于概述共指而言，使用代词指代先行词（小句），比使用短语指代先行词（小句）更为常用。


(c) 2024 中国中文信息学会计算语言学专业委员会

| 共指关系 | 总频次 | 平均频次/链 | 方差 | 比例（%） |
|---|---|---|---|---|
| :homo | 1495.00 | 1.89 | 6.88 | 43.46 |
| :root | 792.00 | 1.00 | 0.00 | 23.02 |
| :reduce | 338.00 | 0.43 | 1.77 | 9.83 |
| :add | 309.00 | 0.39 | 1.09 | 8.98 |
| :alias | 278.00 | 0.35 | 0.75 | 8.08 |
| :pro | 200.00 | 0.25 | 0.91 | 5.81 |
| :encap_pro | 26.00 | 0.03 | 0.01 | 0.76 |
| :encap_alias | 22.00 | 0.03 | 0.02 | 0.64 |
| :illustrate_of_alias | 12.00 | 0.02 | 0.03 | 0.35 |
| :illustrate | 3.00 | 0.00 | 0.00 | 0.09 |
| :illustrate_of_pro | 1.00 | 0.13 | 0.13 | 0.03 |
| :encap_of_alias | 0.00 | 0.00 | 0.00 | 0.00 |
| :encap_of_pro | 0.00 | 0.00 | 0.00 | 0.00 |

表 6: 共指标签频次表

# 6 篇章共指语义图的构建

## 6.1 篇章共指语义图组成部分

本文所构建的篇章共指语义图为单根有向无环图，图中结点为共指元素，边表示共指元素之间的关系。其中，横向关系表示共指元素的句内语义关系，纵向关系表示共指元素的共指关系（见表 4、表 5）。上文中介绍到句子级 CAMR 共有 5 种核心语义角色关系和 47 种非核心语义角色关系，在篇章共指中常用的 CAMR 语义角色关系主要有以下几类。

### 6.1.1 核心语义角色关系

核心语义角色关系沿用 OntoNotes 体系的核心语义角色关系，标签及其代表关系如表 7 所示。

| 核心语义角色关系关系标签 | 关系 |
|---|---|
| arg0 | 原型施事 |
| arg1 | 原型受事 |
| arg2 | 间接宾语、受益者、工具等 |
| arg3 | 出发点、受益者等 |
| arg4 | 终点 |

表 7: 核心语义角色关系汇总

### 6.1.2 非核心语义角色关系

CAMR 中，共有 47 种非核心语义关系，理论上均可以在共指标注中使用，但常见的语义角色关系有以下几种（见表 8）。

## 6.2 篇章共指语义图构建示例

篇章共指语义图为单根有向无环图，图中结点为篇章中所出现的共指概念，边表示共指概念之间的关系，即横向的句内语义关系和纵向的共指关系。本文将从横纵角度探讨篇章共指语义图结构，以证明基于 CAMR 体系标注和构建共指语义结构，有利于揭示篇章结构中共指链的动态变化，挖掘篇章共指语义信息。

我们选用了 CTB8.0 中第 0226 篇新闻，该文本中含有实体共指链 13 条，共指链长度为 1-7 句不等，跨句情况为 1-9 句不等，共出现 7 次句内共指和 47 次句间共指。我们依据共指关系类


(c) 2024 中国中文信息学会计算语言学专业委员会

| 序号 | 非核心语义角色关系 | 中文说明 | 序号 | 非核心语义角色关系 | 中文说明 |
|---|---|---|---|---|---|
| 1 | location | 处所 | 6 | poss | 领属 |
| 2 | manner | 方式 | 7 | subset | 子集 |
| 3 | medium | 媒介 | 8 | superset | 父集 |
| 4 | mod | 修饰 | 9 | topic | 话题 |
| 5 | part-of | 部分 | | | |

表 8: 用于共指标注的常见非核心语义角色关系

型与句内语义关系，架构该文本的篇章共指语义图，如图 1 所示[1]。图 1 以句子为基本单元，每一个虚线框都代表一个句子平面，框的左上部为文本原句，虚线框内不同颜色的实线框为该文本中的共指元素，已用颜色高亮加以区别，且标注在文本原句中。图中的有向实线分为横纵两个方向，横向表示共指元素的句内语义关系（如“:arg0”标签），纵向表示共指元素之间的共指关系（如“:homo”标签）。下面将以该文本为示例，从横纵角度探讨篇章共指语义图结构。

（1）横向：揭示句子内部共指信息。

横向以句子为基本单元，可以揭示：（1）每句共指元素的数量及嵌套情况。如 S1 句共分布有 5 种共指现象，其中有 3 种存在层层嵌套关系。（2）每句的句内共指情况。如 S3 中“两军两国之间的合作”和“上述问题”具有概念同一性，为句内共指。（3）共指元素的句内语义关系。如在 S1 中，“迟浩田”的句内语义关系为“:arg0（施事）”，“坦桑”的句内语义关系为“:mod（修饰）”。

（2）纵向：揭示句子之间概念转移。

纵向以共指链为基本单元，可以揭示：（1）共指链的起止、长度及其分布。如“会谈”共指链，起于 S1，止于 S9，跨句情况为 9 句；共含有 3 个共指元素，分布于 S1、S3、S9，尽管出现次数较少，但几乎贯穿本篇始终。（2）共指元素之间的共指关系。如“会谈”共指链，在全文中并未出现变化；如“两国”共指链在篇章中也会使用“中国和坦桑”这类表达。

总之，语义关系将共指同一句、不同共指链之间的共指元素连结，共指关系将不同句、同一共指链之间的共指元素，构成了共指语义结构的基本骨架。该图在句子层揭示了共指元素的句内语义关系，在篇章级揭示了共指元素之间的共指关系，通过横纵关系综合观察，可总结篇章结构中共指变化的动态性。

## 7 结论与未来工作

共指是近年篇章研究热点，尽管计算语言学领域相关文章更新迭代较快，但其大多为指代消解模型的构建更新。且近年发布的多语言共指资源大多依据 2013 年发布的 OntoNotes 标注体系而构建，在标注体系的根本设定上并未有较大突破。因此，这些语料库均存在共指关系粗疏以及基本不考虑整句语义表示的融合等问题。本文在句子级 CAMR 体系基础之上探索共指标注体系，从整句语义结构探索共指的演变推进；区分了 8 种共指关系类型，对共指关系作出一定阐释；从 CAMR 中继承的句内语义关系，可结合篇章共指关系构建篇章共指语义图。基于该标注体系构建了 100 篇新闻共指语料库。该语料库的标注信息不仅涵盖了频次、位置、间隔距离、句法成分等经典计量指标，还拓展了基于语义结构的计量信息。在未来，我们将从以下三个方面开展工作：首先，在标注体系与标注规模上，我们将扩大新闻语料的标注规模，并尝试探索记叙文、小说等共指现象丰富的语料，完善共指标注方案；其次当标注体系稳定之后，将依据该规范扩大语料库规模。最后，基于共指标注体系发展实体间关系和篇章修辞结构，尝试建构整全的篇章语义结构图。

[1]清晰版可见网页：https://www.camrp.tech/DAMR/


(c) 2024 中国中文信息学会计算语言学专业委员会

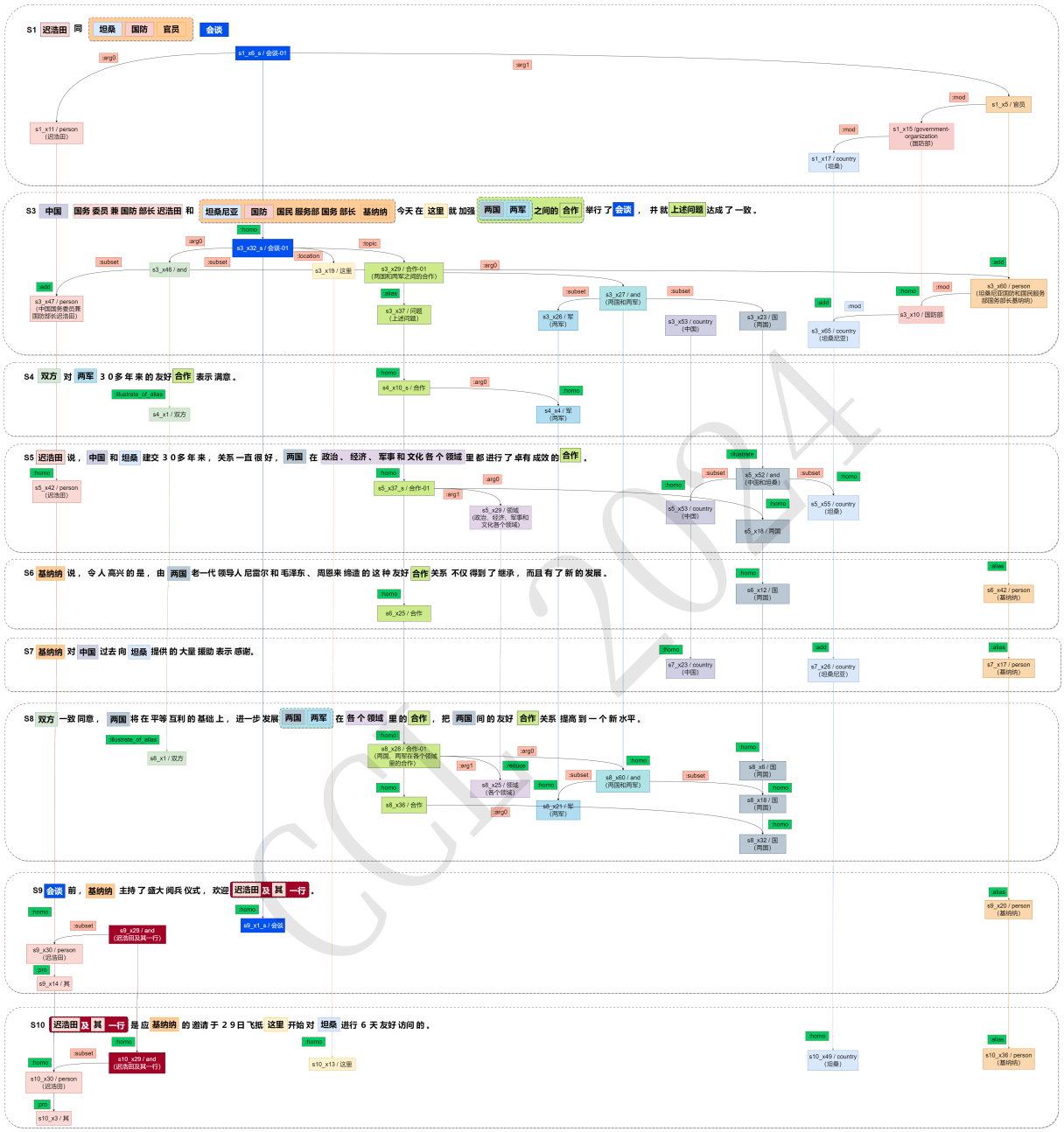



图 1: 篇章共指语义图示例

CCL 2024

## 参考文献

Mira Ariel. 1990. Accessing noun-phrase antecedents. vol. 96. londres: Routlege. *Linguistics*, pages 113–118.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, pages 178–186.

Erik Cambria, Qian Liu, Sergio Decherchi, Frank Xing, and Kenneth Kwok. 2022. Senticnet 7: A commonsense-based neurosymbolic ai framework for explainable sentiment analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3829–3839.

Hong Chen, Zhenhua Fan, Hao Lu, Alan L Yuille, and Shu Rong. 2018. Preco: A large-scale dataset in preschool vocabulary for coreference resolution. *arXiv preprint arXiv:1810.09807*.

Loic De Langhe, Orphée De Clercq, and Veronique Hoste. 2023. Constructing a cross-document event coreference corpus for dutch. *Language Resources and Evaluation*, 57(2):819–848.

Vladimir Dobrovolskii, Mariia Michurina, and Alexandra Ivoylova. 2022. Rucoco: a new russian corpus with coreference annotation. *arXiv preprint arXiv:2206.04925*.

George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie M Strassel, and Ralph M Weischedel. 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. In *Lrec*, volume 2, pages 837–840. Lisbon.

Feride Erkü and Jeanette K Gundel. 1987. The pragmatics of indirect anaphors. In *The pragmatic perspective*, pages 533–545. John Benjamins BV.

Abbas Ghaddar and Philippe Langlais. 2016. Wikicoref: An english coreference-annotated corpus of wikipedia articles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 136–142.

Talmy Givón. 1983. Topic continuity in discourse. *Topic continuity in discourse*, pages 1–498.

Lynette Hirschman and Nancy Chinchor. 1998. Appendix f: Muc-7 coreference task definition (version 3.0). In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998*.

C-T James Huang. 1984. On the distribution and reference of empty pronouns. *Linguistic inquiry*, pages 531–574.

Jun Lang, Bing Qin, Ting Liu, and Sheng Li. 2007. 篇章共指消解研究综述. *Journal of Chinese Language and Computing*, 17(4):227–253.

Ekaterina Lapshinova-Koltunski and Pedro Augusto Ferreira. 2022. *ParCorFull2. 0: A parallel corpus annotated with full coreference*. Saarländische Universitäts-und Landesbibliothek.

Bin Li, Yuan Wen, Li Song, Weiguang Qu, and Nianwen Xue. 2019. Building a chinese amr bank with concept and relation alignments. *Linguistic issues in language technology*, 18.

Ruicheng Liu, Rui Mao, Anh Tuan Luu, and Erik Cambria. 2023. A brief survey on recent advances in coreference resolution. *Artificial Intelligence Review*, 56(12):14439–14481.

Rui Mao, Chenghua Lin, and Frank Guerin. 2018. Word embedding and wordnet based metaphor identification and interpretation. In *Proceedings of the 56th annual meeting of the association for computational linguistics*. Association for Computational Linguistics (ACL).

Rui Mao, Xiao Li, Mengshi Ge, and Erik Cambria. 2022. Metapro: A computational metaphor processing model for text pre-processing. *Information Fusion*, 86:30–43.

Katja Markert, Yufang Hou, and Michael Strube. 2012. Collective classification for fine-grained information status. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 795–804.


(c) 2024 中国中文信息学会计算语言学专业委员会

Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolářová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, et al. 2006. Annotation on the tectogrammatical level in the prague dependency treebank. *annotation manual. Technical Report*, 30:5–11.

Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtskỳ, Amir Zeldes, and Daniel Zeman. 2022. Corefud 1.0: Coreference meets universal dependencies. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4859–4872.

Tim O' Gorman, Michael Regan, Kira Griffitt, Ulf Hermjakob, Kevin Knight, and Martha Palmer. 2018. Amr beyond the sentence: the multi-sentence amr corpus. In *Proceedings of the 27th international conference on computational linguistics*, pages 3693–3702.

Massimo Poesio, Ron Artstein, et al. 2008. Anaphoric annotation in the arrau corpus. In *LREC*.

Massimo Poesio, Jon Chamberlain, Silviu Paun, Juntao Yu, Alexandra Uma, and Udo Kruschwitz. 2019. A crowdsourced corpus of multiple judgments and disagreement on anaphoric interpretation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1778–1789.

Ina Rösiger. 2016. Scicorp: A corpus of english scientific articles annotated for information status analysis. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1743–1749.

Ina Rösiger. 2018. Bashi: A corpus of wall street journal articles annotated with bridging links. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Olga Uryupina, Ron Artstein, Antonella Bristot, Federica Cavicchio, Kepa Rodriguez, and Massimo Poesio. 2016. Arrau: Linguistically-motivated annotation of anaphoric descriptions. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2058–2062.

Renata Vieira and Massimo Poesio. 2000. An empirically based system for processing definite descriptions. *Computational Linguistics*, 26(4):539–593.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. 2013. Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA*, 23:170.

冉晨. 2024. 现代汉语中数量名回指语的指称性质与回指确认方式. 语言教学与研究, pages 59–68.

刘东虹. 2014. 书面语篇中的抽象实体回指研究. 华中师范大学出版社.

吴丹苹. 2021. 基于语料库的现代汉语概述回指研究. Ph.D. thesis, 博士学位论文]. 杭州: 浙江大学.

周炫余, 刘娟, and 卢笑. 2014. 篇章中指代消解研究综述. 武汉大学学报: 理学版, (1):24–36.

孔芳, 葛海柱, and 周国栋. 2021. 篇章视角的汉语零指代语料库构建. 软件学报, 32(12):3782–3801.

宋宏. 2010. 人称代词语篇回指研究.

宋洋 and 王厚峰. 2015. 共指消解研究方法综述. 中文信息学报, 29:1–12.

寇鑫 and 徐坤宇. 2023. 抽象回指的指称内容与可及性研究——以“这”和“这件事”为例. 语言教学与研究, pages 88–100.

徐赳赳. 2003. 现代汉语篇章回指研究. 中国社会科学出版社.

曹娜 and 曹贤文. 2024. 汉语二语学习者主语回指语显隐的多因素分析. 语言教学与研究, pages 24–34.

李榕 and 王元鑫. 2021. 中高级阶段韩国留学生汉语篇章第三人称回指的习得研究. 世界汉语教学, 35:276–288.

杨永生 and 肖奚强. 2020. 韩国学生汉语“这/那”句习得考察. 华文教学与研究, (1):69–75.

王军. 2004. 英语叙事篇章中间接回指释义的认知研究. 苏州大学出版社.


(c) 2024 中国中文信息学会计算语言学专业委员会

王军. 2013. 英汉语篇间接回指. 商务印书馆.

王秀丽. 2012. 篇章分析中的概述回指. 当代语言学, 14(3):301–306.

石艳华. 2014. 认知激活框架下的汉语篇章回指研究.

胡霄钦 and 王秀丽. 2021. 法汉指称链条平行语料库的建设与应用. 语料库语言学, 8:112–121.

蒋平. 2021. 存现宾语的零形回指及其认知语言学解释. 山东外语教学, 42:22–30.

许余龙. 2004. 篇章回指的功能语用探索*:* 一项基于汉语民间故事和报刊语料的研究. 上海外语教育出版社.

陈远哲, 匡俊, 刘婷婷, 高明, 周傲英, et al. 2019. 共指消解技术综述. 华东师范大学学报 *(*自然科学版*)*, 5:16–35.

CCL 2024

# 汉语中介语词同现网络研究

**钱隆[1]，赵慧周[2]，丁芊[3]，王治敏[4]***
1.北京语言大学 国际中文教育研究院/北京
2. 北京语言大学 信息科学院/北京
3. 安徽信息工程学院 通识教育与外国语学院/安徽芜湖
4.广东外语外贸大学 中国语言文化学院/广东广州
qianlong_1221@163.com zhaohuizhou@blcu.edu.cn
dingqian0801@163.com wangzm000@qq.com

## 摘要

近年来，运用复杂网络方法进行语言学研究已成为数字人文研究的一条新路径。本文基于214篇日本汉语学习者的书面作文，构建了6个不同能力水平的汉语中介语词同现网络，并探讨了这些网络的结构特性及其动态演变过程。研究结果显示，所有的汉语中介语词同现网络均呈现出小世界属性、无标度属性、异配性和层级结构等复杂网络的特性。这些特性揭示了汉语学习者在词汇使用方面的特定模式：低水平学习者更倾向于将低频词汇与高频词汇进行连接，这可能与学习者减轻认知负荷的习得模式有关；学习者语言水平的提升，中介语网络参数会逐渐向母语者靠拢，但是无法达到母语者的水平；此外，本研究还观察到，语言错误会对中介语网络结构产生影响，引起网络结构的变异。

**关键词：** 复杂网络；汉语中介语；小世界属性；无标度属性；异配性

# A Study on Chinese Interlanguage Co-occurrence Networks

**QIAN Long[1] ZHAO Huizhou[2] DING Qian[3] WANG Zhimin[4]**
1. Research Institute of International Chinese Language Education,
Beijing Language and Culture University / Beijing
2. College of Information Science,
Beijing Language and Culture University / Beijing
3. Faculty of General Education and Foreign Languages,
Anhui Institute of Information Technology / Anhui Wuhu
4.Faculty of Chinese Language and Culture,
Guangdong University of Foreign Studies / Guangdong Guangzhou

## Abstract

Within the framework of complex network theory, this paper constructs 6 Chinese interlanguage co-occurrence networks based on 214 written compositions by Japanese L2 learners of Chinese, categorized by different proficiency levels. It explores the structural characteristics and dynamic evolution of these networks. The results reveal that all the Chinese interlanguage co-occurrence networks exhibit complex network properties such as small-worldness, scale-freeness, disassortativity, and hierarchical structure. These characteristics highlight specific patterns in vocabulary usage among Chinese learners: learners at lower levels tend to connect new words with high-frequency words, possibly related to the anchoring effect and the principle of least effort; as learners' proficiency

*通讯作者 Corresponding Author

**基金：**本文系2018 年国家社会科学基金重大项目"基于'互联网+'的国际汉语教学资源与智慧教育平台研究"（18ZDA295）的研究成果。

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

increases, the parameters of the interlanguage networks gradually converge towards those of native speakers, yet they do not reach the native speakers' level; moreover, this study observes that language errors impact the interlanguage network structure, causing variations.

**Keywords:** Complex Networks , Co-occurrence , Chinese Interlanguage , Small-World Property , Scale-Free Property , Disassortativity

## 1 引言

在自然界和人类社会中，网络现象无处不在。网络是可以用图来表示的系统，系统中的元素是网络的节点，而元素之间的关系则是网络的边(Barabási, 2002; Watts, 2003)。绝大多数现实世界的网络，例如：社交网络(Davidsen et al., 2002)、学术引文网络(Greenberg, 2009)、生物基因网络(Levine and Davidson, 2005)、流行病传播网络(Danon et al., 2011)和电力网络(Pagani and Aiello, 2013)等都属于复杂网络范畴，它们既不是完全规则的，也不是完全随机的(Watts and Strogatz, 1998)。自Watts and Strogatz(1998)和Barabási and Albert(1999)在复杂网络研究中分别引入了"小世界网络"和"无标度网络"的概念以来，复杂网络理论便迎来了快速发展的阶段。此后，广泛的研究表明，小世界属性和无标度分布在自然界和人类社会中均普遍存在。这些研究成果不仅进一步丰富了复杂网络的理论，也为跨学科研究提供了新的视角与方法。

语言普遍被视为一种复杂适应系统(Briscoe, 1998; Ellis and Larsen-Freeman, 2006; Ellis and Larsen-Freeman, 2009; Cameron and Larsen-Freeman, 2007)。这使得复杂网络分析方法应用于语言学研究成为一种可能。复杂网络方法能够捕获网络的整体特征，揭示相变，并为系统分析提供有效的定量测量(Cong and Liu, 2014)。在语音、词汇、句法、语义等不同层面上，自然语言都展现出复杂网络的特性(刘海涛, 2022)。研究发现，汉字网络(Li and Zhou, 2007)、句法依存网络(Cancho et al., 2004)、语义网络(Borge-Holthoefer and Arenas, 2010)等均具有复杂网络特性。随着复杂网络理论的不断深化，研究者开始关注于中介语网络的结构特性以及这些特性在语言习得和应用中的影响。

中介语是指第二语言学习者在其母语（第一语言）和目的语（第二语言）之间形成的一种心理语言系统，其特点包括动态性、系统性、过渡性和创造性(Selinker, 1972)。近年来，随着复杂网络方法的应用与发展，越来越多的研究聚焦于中介语网络的结构特性及其发展过程。例如：Jiang et al(2019)在研究英语中介语句法依存网络时指出，二语学习者的句法发展并不会像一语学习者那样出现"句法涌现"现象。Hao et al(2021)揭示了复杂网络参数能有效区分二语学习者的语言熟练度。韩笑等(2021)发现中介语句法复杂度网络的演变主要体现在网络强度及中心节点的变化上。王士丛(2021)则观察到汉语中介语网络的国别化差异，不同国别学习者词汇网络内部存在致密性差异。虽然这些研究丰富了我们对中介语网络结构与动态性的认识，但依然存在一些不足。首先，大部分研究集中于探讨中介语网络的总体特性及其演变，却往往未能深究这些特征以及变化背后的具体成因。其次，当前的中介语网络研究主要集中在句法依存网络的构建上，而较少使用词同现网络方法，尽管后者在直观反映词汇间的搭配关系和结构特征方面具有明显优势(Liu and Cong, 2013)。最后，二语学习者的语言输出会产生较多的语言错误，这些错误是中介语的重要组成部分，然而现有的研究对此并未给予足够的关注和讨论。

鉴于此，本文通过构建与分析汉语中介语词同现网络，旨在揭示汉语作为第二语言学习者在语言习得过程中的模式与特征，深入分析这些模式和特征背后的成因，并考察语言错误对中介语网络的具体影响，进而为汉语教学提供理论依据与实践指导。

## 2 研究设计

### 2.1 语料

本研究使用的语料分为中介语语料和母语者语料两部分。中介语语料采集自全球汉语中介语语料库（简称"全球库"）[-1]。全球库是一个大型的汉语中介语语料库，迄今为止已累积收集了约2367万字的中介语资料(张宝林、崔希亮, 2022)。本研究选取了全球库中1530篇日本汉

[-1]全球汉语中介语语料库网址：http://qqk.blcu.edu.cn/


(c) 2024 中国中文信息学会计算语言学专业委员会

语学习者的书面作文，总字数约为41.7万字。母语者语料则采集自作文网[0]，这是一个开放的中国中小学生作文投稿平台，涵盖了广泛的年级和题材。本研究从该平台收集了60篇中学生的期末考试作文，总计3.9万字。语料采集完毕后，依据《欧洲语言共同框架》（CEFR）中的整体书面产出（Overall Written Production）和写作评定网格（Written Assessment Grid）对1530篇作文进行分级。《欧洲语言共同框架》将语言使用者的语言水平分为六个等级，分别是：A1、A2级表示基础使用者（Basic user），B1、B2级表示独立使用者（Independent user），C1、C2级表示熟练使用者（Proficient user）(of Europe, 2018)。分级工作由三名评分员和一名专家共同完成，三名评分员均拥有超过三年的评分经验，而专家则多次担任国家级汉语水平考试的阅卷组长。具体步骤如下：首先，每位评分员独立地对每篇作文按照A1至C2的等级进行分级。若三位评分员给出的分级完全一致，或者其中两位的评分一致且第三位的评分与此一致评分相距一个等级，则认定该分级结果有效，据此确定作文的最终等级。如果三位评分员的评分全都不一致，或两位一致但第三位的评分与此一致评分相距两个等级或以上，则认定该分级结果无效，需提交给专家进行重新分级。最终，有效分级的作文数量为1231篇，无效分级的作文数量为299篇，有效分级的比率达到了80.46%。为验证分级结果的一致性，采用SPSS 27.0软件进行了肯德尔和谐系数（Kendall's coefficient of concordance）检验，具体检验结果展示于表1中。

| 评分员数 | 肯德尔$W^a$ | 卡方 | 自由度 | 渐近显著性 |
|---|---|---|---|---|
| 3 | 0.747 | 3425.435 | 1529 | 0.000 |

Table 1: 分级数据肯德尔和谐系数检验结果

由表1可知，肯德尔和谐系数检验呈现出显著性(p=0.000<0.05)，三位评分员的分级结果具有一致性。同时，肯德尔和谐系数为0.747，介于0.6-0.8之间，说明三位评分员的分级一致性较强，分级结果可以采纳。结合有效的分级结果和专家的分级结果，本研究对1530篇语料全部进行了分级。

中介语语料分级结束后，使用Jieba软件对母语语料和中介语语料进行了分词处理，并人工修改了部分错误。母语语料与中介语语料详细信息见表2。

| 子库 | A1 | A2 | B1 | B2 | C1 | C2 | 母语 |
|---|---|---|---|---|---|---|---|
| 篇数 | 94 | 410 | 582 | 290 | 115 | 39 | 60 |
| 字数 | 11828 | 79943 | 158241 | 92665 | 44626 | 23218 | 39415 |
| 词数 | 10526 | 67585 | 127461 | 72837 | 34672 | 17597 | 25227 |

Table 2: 母语语料与中介语语料详细信息

考虑到六个等级的中介语语料与母语语料总量不等，为避免语料规模差异对研究结果产生影响，本研究对各子库内的语料进行了平衡处理。具体措施包括：首先，从每个等级的作文中剔除词数的极端值，以保证各等级学习者的写作水平具有一致性；随后，通过简单随机抽样的方式调整每个子库的规模，确保其词数控制在5000±50词的范围内，以实现语料库之间的容量均衡。最终我们抽取了214篇中介语作文以及12篇母语作文。语料抽样信息见下表。

| 子库 | A1 | A2 | B1 | B2 | C1 | C2 | 母语 |
|---|---|---|---|---|---|---|---|
| 篇数 | 69 | 52 | 31 | 26 | 21 | 15 | 12 |
| 字数 | 17627 | 17827 | 17785 | 17881 | 18140 | 18111 | 17893 |
| 词数 | 5036 | 5014 | 5018 | 4957 | 5020 | 5021 | 4981 |

Table 3: 母语语料与中介语语料抽样信息

[0]作文网网址：https://www.zuowen.com/


(c) 2024 中国中文信息学会计算语言学专业委员会

## 2.2 网络构建

无论网络如何复杂，其构成的基本要素都很简单。网络构成的基本要素就是节点（vertices）和边（edges）(Newman, 2002)。在语言网络中，节点通常由各类语言单位充当，如：词、音素、语素、汉字、汉字部件等；而边则表示节点与节点之间的关系(Liu, 2008; Cong and Liu, 2014)。

在构造以词为节点的网络时，采用词同现的方式来构建语言网络是一种常用方法。词同现揭示了语言的基本属性之一——线条性，即词汇在文本中按照先后顺序线性排列。词同现网络以词（词型，word types）为节点，两个同时出现（相邻）的词通过边相连(Liang et al., 2009)。词同现关系存在有向和无向之分。尽管有向网络能够展示词汇序列的方向性，但是无向网络则更适合于分析词汇同现的结构及其动态变化(Garg and Kumar, 2018)。因此，本研究采用无向词同现网络的构拟方法来构建语言网络，即从语料库中提取所有不同的二元词同现对，见图1（a），并将其转换成词同现网络，见图1（b）。




(a) 词同现分析　　(b) 词同现网络

Figure 1: 例句的词同现分析和词同现网络

在构建网络之初，本研究首先对语料进行了处理，去除了所有的非完句标点符号（“，”“：”“；”等）和外文符号。接着，参考刘海涛(2022)的做法，从各子库中提取二元词同现对，直到遇到完句符号（“。”“！”“？”），此时跳转至下一句继续提取，确保二元词同现对不跨句提取。随后，将提取的二元词同现对依次放入Excel表格中。最后，使用Createpajek软件，将Excel中的二元词同现对转化为词同现网络。据此，本研究成功构建了六个不同等级（A1、A2、B1、B2、C1、C2）的中介语网络以及一个母语网络。

为了探究中介语网络是否具有小世界属性，本研究进一步构建了六个相应的随机网络进行比较分析，这些随机网络与对应的中介语网络具有相同的节点数、边数和平均度。此外，为了观察语言错误对中介语网络结构的影响，本研究还对六个中介语网络中的语言错误进行了人工修正，创建了六个无语言错误的中介语词同现网络。构建完成所有网络后，使用Pajek软件对这些网络的特征进行了数据提取。

## 2.3 网络的统计特征

### 2.3.1 网络密度

网络密度$D$指的是网络中实际存在的边（即连接）数与网络中可能存在的最大边数之间的比例。该指标可以描述网络中节点间连接的紧密程度(Wasserman and Faust, 1994; Scott, 2000)，公式如下：

$$D = \frac{2M}{N(N-1)}$$

其中，$M$是网络中实际边的数量，$N$是网络的节点数。网络密度的值范围从0到1。0表示一个完全无连接的网络（即没有任何边），而1表示一个完全连接的网络（即网络中任意两个不同的节点之间都存在边），通常真实网络的密度总是小于0.5(Mayhew and Levinger, 1976)。高密度值意味着网络中的节点之间存在较多的连接，表明词与词之间的连接很紧密；相反，低密度值表明网络中的连接较少，词与词之间的连接较为稀疏。

第二十三届中国计算语言学大会论文集，第712页-第725页，太原，中国，2024年7月25日至28日。
卷1：主会论文
(c) 2024 中国中文信息学会计算语言学专业委员会

### 2.3.2 度分布

假设一个有$N$个节点的网络中，边都是无向且无权的，那么节点$i$具有的边的数量则是该节点的度$k_i$。节点的度分布$p(k)$指的是网络中随机选择的节点恰好为$k$度的概率，如果度分布$p(k)$满足幂律分布($p(k) \propto k^{-\gamma}$，$\gamma$为常数)，则网络被称为无标度网络(scale-free network)。通常，真实网络的度分布都服从幂律分布，而随机网络都服从泊松分布(Barabási and Albert, 1999)。

### 2.3.3 平均路径长度

在无向无权的网络中，两个节点之间的距离$d$为连接二者的最短路径的边的数目(刘涛et al., 2005)，平均路径长度$L$则是网络中任意两个节点之间距离的平均值。

$$L = \frac{2\sum_{i<j} d_{ij}}{N(N-1)}$$

其中，$d_{ij}$表示节点$i$和节点$j$之间的最短距离，$N$为网络的节点数。平均路径长度短表示网络中任意两个节点之间可以通过较少的中间节点相互到达，网络具有较高的全局连通性；反之，则表示网络的连通性较差。平均路径长度是衡量网络是否具有小世界属性的重要指标之一(Albert and Barabási, 2002)，绝大多数真实网络的平均路径长度要小于随机网络(Watts and Strogatz, 1998)。

### 2.3.4 聚集系数和度相关平均聚集系数

聚集系数$C$是用来描述网络中节点之间聚集程度的系数。聚集系数可以反映连接到同一个节点上的两个节点相互连接的概率(Newman, 2010)。假设节点$i$通过$k_i$条边和其他的$k_i$个节点相连，那么这$k_i$个节点都可以视为节点$i$的邻接点。如果$i$的所有邻接点之间又都互相连接，那么他们之间存在$k_i(k_i\text{-}1)/2$条边。如果将$M_i$看作是$i$的所有邻接点之间存在的实际边数，那么$M_i$与$k_i(k_i\text{-}1)/2$的比值则是节点$i$的聚集系数$C_i$(刘涛et al., 2005)。

$$C_i = \frac{2M_i}{k_i(k_i-1)}$$

一个网络的聚集系数$C$则是整个网络中所有节点聚集系数的平均值(**?**)。聚集系数的值介于0和1之间。0表示网络中没有节点形成紧密的群组，而1表示网络是完全聚集的，即每个节点的所有邻居节点都彼此直接相连。聚集系数也是小世界效应的重要衡量标准之一(Albert and Barabási, 2002)。通常，在相同节点数和边数的情况下，真实网络的聚集系数要远大于随机网络的聚集系数(Watts and Strogatz, 1998)。

度相关聚集系数$C(k)$（Degree-dependent clustering coefficient）则是聚集系数的扩展，指的是所有度数为$k$的节点的平均聚集系数。在研究中，可以通过该指标分析网络的度分布和聚集系数的关系，以揭示网络的层级结构。当度相关聚集系数$C(k)$遵循幂律分布（$C(k) \propto k^{-\gamma}$，$\gamma$为常数），这表明网络往往呈现出层级化的拓扑结构特征(Ravasz and Barabási, 2003)。这种层级结构可被视为网络的一种组织模式(刘海涛, 2017)。幂律分布意味着高度节点较少参与形成三角结构（三元闭包），通常都是低度节点倾向形成三元闭包构成较为紧致的子网络，子网络间的连接相对较少，而高度节点则将这些子网络连接成一个连通的网络。

### 2.3.5 同配系数

同配系数（assortativity coefficients）是用来量化网络中节点的连接倾向的一个指标，可以用来反映网络中节点连接的同质性。基于节点度的同配系数可以衡量高度节点是否倾向于与其他高度节点相连，以及低度节点是否倾向于与其他低度节点相连。公式如下：

$$r = \frac{M^{-1}\sum_i j_i k_i - \left[M^{-1}\sum_i \frac{1}{2}(j_i + k_i)\right]^2}{M^{-1}\sum_i \frac{1}{2}(j_i^2 + k_i^2) - \left[M^{-1}\sum_i \frac{1}{2}(j_i + k_i)\right]^2}$$

其中，$M$是边的总数，$j_i$和$k_i$是连接到第i条边两端节点的度。同配系数的取值范围在-1和1之间，当同配系数为正值（r >0）时，表示网络具有同配性，即类似度数的节点更可能相连；同


(c) 2024 中国中文信息学会计算语言学专业委员会

配系数为负值（r <0）时，网络具有异配性，即高度数的节点倾向于与低度数的节点相连；同配系数接近0意味着网络中的连接模式更接近随机，没有明显倾向于同配性或异配性(Newman, 2002)。真实世界的许多网络存在同配现象(Newman, 2003a)；而大多数语言网络则呈现出异配性(刘海涛, 2017)。

## 3 结果

### 3.1 网络基本概貌

根据上文所述的网络构建方法，绘制了6个水平的汉语中介语词同现网络图，见图2。

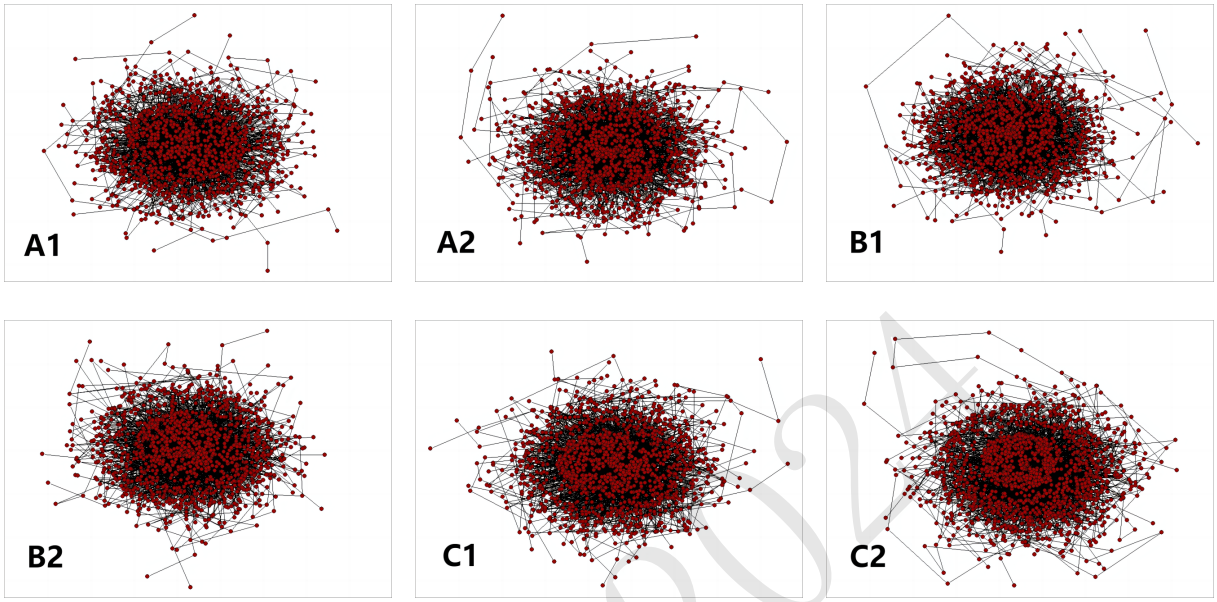



Figure 2: 汉语中介语词同现网络图

图2中的点表示网络的节点，即词型；节点与节点之间的线则表示网络的边，即两个词之间的同现关系。通过图2可以直观地观察到中介语词同现网络的复杂性。然而，由于节点和边的数量庞大，直接从图示角度分析网络结构变得相当困难。因此，必须借助相应的网络数据进行深入的分析。

表4提供了6个汉语中介语词同现网络和汉语母语词同现网络的基本信息，包括网络的节点数、边数、词数、型例比（TTR）和密度。其中，节点数就是网络的词型数（word types），词数表示网络的词例数（word tokens）。

| 网络 | A1 | A2 | B1 | B2 | C1 | C2 | 母语 |
|---|---|---|---|---|---|---|---|
| 节点数 | 1193 | 1244 | 1225 | 1290 | 1342 | 1563 | 1926 |
| 边 | 3518 | 3631 | 3830 | 3876 | 3938 | 4069 | 4262 |
| 词数 | 5020 | 5014 | 5018 | 4987 | 5048 | 5021 | 4981 |
| TTR | 0.238 | 0.248 | 0.244 | 0.259 | 0.266 | 0.311 | 0.387 |
| 网络密度 | 0.0049 | 0.0047 | 0.0051 | 0.0047 | 0.0044 | 0.0033 | 0.0023 |

Table 4: 母语语料与中介语语料抽样信息

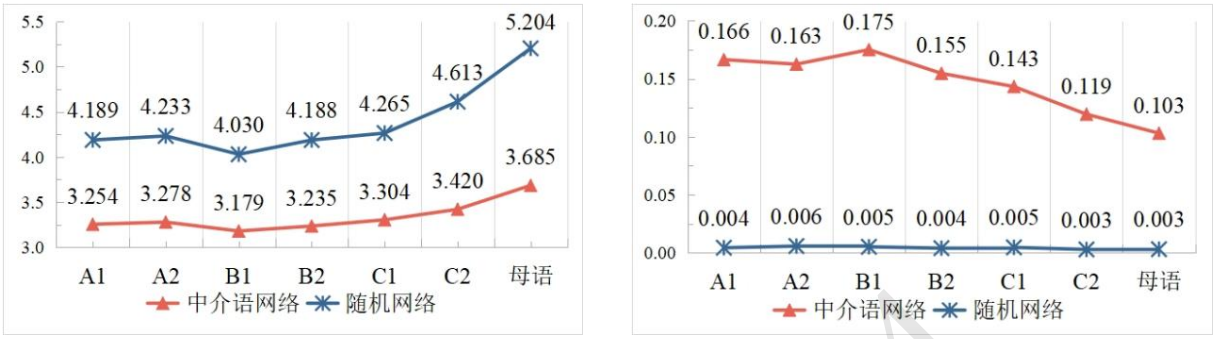
在相同的文本规模下，TTR可以有效衡量文本的词汇丰富程度(Richards, 1987)，由表4可知，随着学习者能力水平的提升，汉语中介语词同现网络的节点数和TTR的值都在逐渐增长。学习者语言水平的提升，书面写作中会使用更加丰富的词汇。不过，这种增长并非线性，节点数和TTR的值都在A1、A2和B1水平下产生了波动，而B1到C2则是线性增长。除此之外，6个水平的网络节点数和TTR的值都未超过母语者网络。


(c) 2024 中国中文信息学会计算语言学专业委员会

网络密度的变化与TTR正好相反，数值随着学习者能力水平的提升而波动下降，中介语网络6个水平的网络密度均高于母语网络。这表明，中介语网络要比母语网络更加紧密，而且从低水平到高水平，网络密度趋于稀疏，并逐渐向母语网络靠拢。

### 3.2 网络的小世界属性

平均路径长度和聚集系数是衡量网络是否具有小世界属性的重要参数。当网络同时具有较高的聚集系数和较短的平均路径长度时，这种网络就是一种小世界网络(Watts and Strogatz, 1998; Albert and Barabási, 2002)。中介语网络与随机网络的平均路径长度和聚集系数详情见图3。




(a) 平均路径长度　　(b) 聚集系数

Figure 3: 小世界属性统计结果

由图3可以清晰的看出，与随机网络相比，同一水平、相同规模下中介语网络都拥有更短的平均路径长度和更大的聚集系数，6个中介语网络均具有小世界属性。小世界属性表明，中介语中的词汇通过同现关系形成了高度结构化的网络。这种结构化表明词与词之间的同现搭配不是随机分布的，而是按照某种内在的逻辑和规律相互关联。

小世界属性揭示了中介语系统内部语言要素是紧密联系的。这种结构可能反映了学习者在二语习得过程中，对词汇和句法结构内在逻辑的适应和理解。Corominas-Murtra et al(2009)曾发现儿童一语者在24个月时，句法图的全局拓扑会从树状结构急剧转变成小世界网络，出现一种“句法涌现”。而Jiang et al(2019)则发现，英语二语学习者句法网络并没有像一语者那样出现“句法涌现”，这表明学习者习得一语句法和习得二语句法的方式是不同的，二语学习者会依靠母语和目的语之间的句法相似性来接近目的语。本研究也有类似地发现，汉语中介语词同现网络中，最低水平（A1）的中介语网络已经出现了小世界属性。即使是在非母语的使用中，低水平的汉语学习者也能有效地构建词汇间有意义的同现搭配，形成一个紧凑而高效的网络。

除此之外，我们还发现平均路径长度和聚集系数的波动与节点数和TTR类似，从B1到C2水平都是是线性增长或下降，而波动主要发生的在A1、A2和B1水平。

### 3.3 网络的无标度属性

网络研究中，随机网络的度分布通常都服从泊松分布（Poisson distribution），而真实网络的度分布通常都服从幂律分布（Power-law distribution），当一种网络的度分布服从幂律分布时，通常把这种网络称作无标度（scale-free）网络(Barabási, 2002)。为了探究汉语中介语词同现网络是否具有无标度属性，本研究提取了6个中介语网络的度数，并在双对数坐标轴中观察度的分布服从何种分布，详情见图4。

由图4可知，6个网络度分布的幂律拟合决定系数$R^2$均在0.9以上，这表明6个中介语网络的度分布均服从幂律分布，汉语中介语词同现网络是无标度网络。无标度网络是一种特殊类型的网络，由于度分布遵循幂律分布，网络中存在少数几个高度连接的节点，大多数节点则只有少数的连接(Barabási and Albert, 1999; Newman, 2003b)，即无论学习者处于低水平还是高水平，中介语网络中始终存在一些高度连接的中心枢纽节点。无标度网络的特性也反映了中介语系统的适应性和潜在的演化路径。随着学习者语言能力的提高，新的词汇可能成为新的中心枢纽节点，原有的网络结构可能发生调整，以适应更复杂的表达需求。

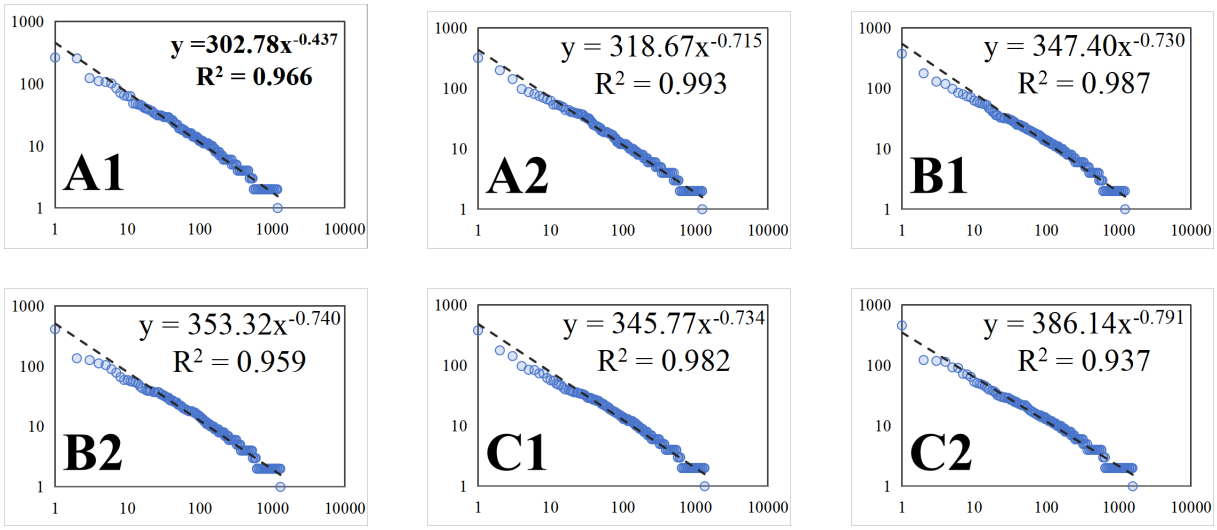



Figure 4: 网络度分布与幂律拟合结果

| | A1 | | A2 | | B1 | | B2 | | C1 | | C2 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 词序 | 节点 | 度数 | 节点 | 度数 | 节点 | 度数 | 节点 | 度数 | 节点 | 度数 | 节点 | 度数 |
| 1 | 我 | 264 | 的 | 321 | 的 | 376 | 的 | 409 | 的 | 380 | 的 | 461 |
| 2 | 的 | 256 | 我 | 200 | 我 | 178 | 我 | 135 | 我 | 178 | 了 | 123 |
| 3 | 是 | 123 | 是 | 143 | 了 | 130 | 是 | 126 | 了 | 143 | 我 | 119 |
| 4 | 了 | 110 | 有 | 98 | 是 | 119 | 了 | 111 | 是 | 98 | 是 | 115 |
| 5 | 她 | 107 | 他 | 87 | 很 | 99 | 有 | 104 | 很 | 85 | 和 | 93 |

Table 5: 母语语料与中介语语料抽样信息

表5展示了6个中介语词同现网络各自度数最高的5个节点。由表5可知，“我”“的”“是”“了”“她”“有”“他”“很”“和”是中介语中最常见的中心枢纽节点。其中，“我”“的”“是”三个节点在6个网络中均出现。不过，中心枢纽节点的位置也并非一成不变。从A1到C2可以明显地观察到代词“我”“她”“他”的地位在下降，助词“的”“了”的地位在上升，其中“的”是A2到C2，5个网络最中心的枢纽节点。

### 3.4 同配系数

同配性和异配性是网络中节点连接的两种不同的方式。一般学者合作网络、电影演员合作网络、公司董事合作网络等都具有同配性，即高度节点和高度节点连接，低度节点和低度节点连接；而因特网、万维网、神经网络等则具有异配性，即高度节点通常和低度节点连接(Newman, 2002)。图5展示了中介语网络和母语网络以及其对应的随机网络的的同配系数。

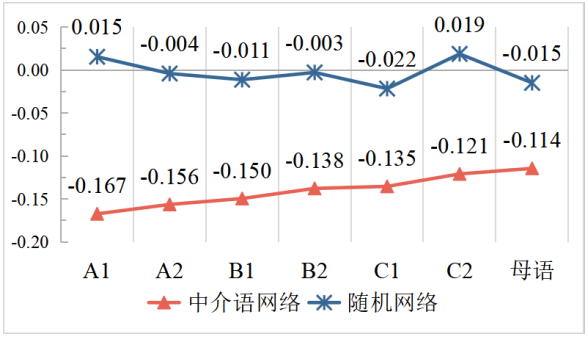



Figure 5: 同配系数


(c) 2024 中国中文信息学会计算语言学专业委员会

由图5分析得知，六个中介语网络的同配系数均呈现负值，揭示出不同层次的汉语中介语词汇共现网络均展现出异配性特征。这表明在中介语网络中，高度节点与低度节点的连接模式更为常见。与此形成对比的是，六个随机网络的同配系数则趋近于0，其值在0的附近波动，表明随机网络中节点的连接方式较为无序，没有表现出明显的同配或异配倾向。值得注意的是，尽管所有中介语网络都显示出异配性，同配系数的数值却随着学习者水平的提高而逐渐上升（绝对值降低），中介语的同配系数逐渐向母语者靠拢。这表明，在初级水平的网络中，异配性特征更加明显，而在高级水平的网络中，这一特征则相对减弱。

### 3.5 层级结构

网络的层级结构（Hierarchical organization）是一种特定的网络拓扑，其中节点按照不同的层次或级别进行组织，形成了明显的上下级关系。当度相关聚集系数$C(k)$满足幂律分布时，网络往往呈现出层级结构。为了探究词同现网络是否呈现层级结构，本研究计算了网络中每一个度数的平均聚集系数，并在双对数坐标轴中观察度相关聚集系数的分布服从何种分布，详情见图6。

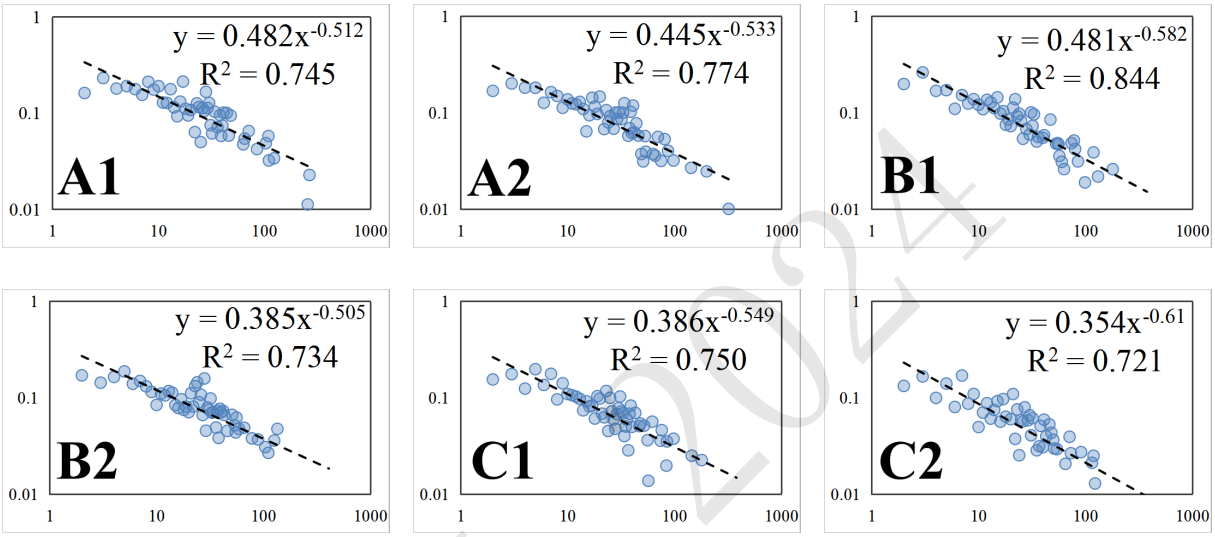



Figure 6: 度相关聚集系数拟合结果

由图6可知，六个网络度相关聚集系数的幂律拟合决定系数$R^2$均超过了0.7，拟合结果在可接受范围内。六个中介语网络均呈现出层级结构。这表明，学习者在语言使用方面即便处于初级水平，词汇之间的搭配也已经较为紧密，呈现出明显的组织和结构。这可能说明学习者在初级水平已经形成对目的语的系统性认识，建立了基本的词汇和语法体系。

### 3.6 纠正语言错误后的中介语网络

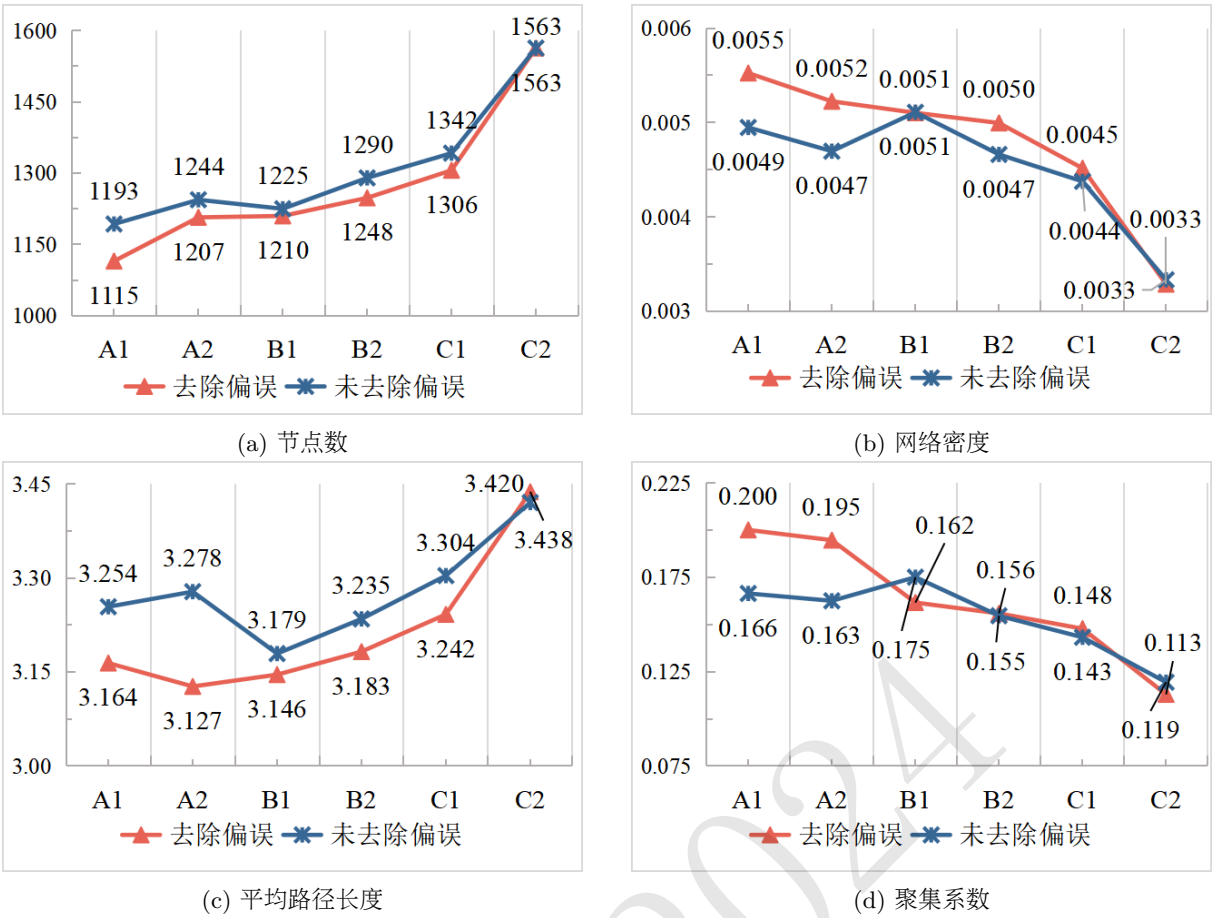
由表4和图3、5可知，六个不同能力水平的网络在节点数、网络密度、平均路径长度以及聚集系数等关键参数上，并未显示出一致的线性变化趋势。在A1、A2和B1等级上，参数数值出现了波动。这种规律性波动可能与学习者在语言使用中的错误有关。鉴于此，本研究对所有网络中的语言错误进行了人工校正。中介语网路中错误数量见表6。

| 网络 | A1 | A2 | B1 | B2 | C1 | C2 |
|---|---|---|---|---|---|---|
| 错误数量 | 549 | 374 | 327 | 235 | 144 | 72 |

Table 6: 中介语网络中语言错误的数量

由表可知，网络规模相同，学习者的能力水平越低，网络中包含的语言错误就越多。为了探究这些语言错误对中介语网络产生了怎样的影响，我们统计了纠正语言错误后的节点数、网络密度、平均路径长度和聚集系数。纠正错语言误后的网络参数见图7。


(c) 2024 中国中文信息学会计算语言学专业委员会




(a) 节点数　(b) 网络密度

(c) 平均路径长度　(d) 聚集系数

Figure 7: 纠正语言错误后的网络参数

由图7可知，经过人工纠正网络中的语言错误后，节点数、网络密度和聚集系数这三项指标的开始表现出线性变化趋势。尽管平均路径长度的数据并未完全遵循线性变化，但在A1、A2、B1水平，这项指标的波动幅度已有所降低。可见，网络中的语言错误实际对网络的变异产生了影响。

## 4　讨论

### 4.1　学习者词汇同现搭配的动态演变

从网络分析的数据中，可以清晰的观察到日本汉语学习者词汇同现搭配的动态演变过程。前文分析中提到，中介语词同现网络表现出小世界属性和异配性特征。从A1到C2水平，中介语网络的平均路径长度增加，聚集系数下降，异配性特征减弱。这三个参数有效反映学习者的写作词汇丰富度和句法水平的演变过程。在A1、A2和B1水平，学习者的词汇量和句法手段都很有限，表达不够灵活自由。为了传达正确的信息，学习者只能在有限的词汇间进行搭配和组合，节点和节点之间就会产生更多的连接，形成紧密连接的群组，即网络中存在许多相互紧密连接的三角结构（三元闭包），从而导致网络具有更大的聚集系数。并且，网络之间有限的搭配和组合也非常依赖高度节点的参与，高度节点通常都是学习者最早掌握并熟知的词，通过迂回的表达使用高度节点，学习者才可以在写作中传达出更多的信息。这种紧密的节点连接和对高度节点的依赖也意味着任意两节点之间可以通过更少的中间节点互相到达，从而导致较低的平均路径长度。随着学习者能力水平的提升，其掌握的词汇越来越多且越来越丰富，使用的句法结构也更加多样，写作表达更加灵活、自由，不再依赖于有限的词汇或固定搭配，尤其是对高度节点的依赖，这使得网络中节点间的连接更加分散，网络的聚集系数逐渐减小，同配系数逐渐上升，以及平均路径长度逐渐增加。这些参数的变化共同揭示了学习者在逐步接近目标语言过程中，从简单到复杂的语言使用发展轨迹，反映了学习者在不断拓展其语言表达的广度和深度。

第二十三届中国计算语言学大会论文集，第712页-第725页，太原，中国，2024年7月25日至28日。
卷1：主会论文
(c) 2024 中国中文信息学会计算语言学专业委员会

### 4.2 学习者语言习得的认知机制

根据前文的分析可知，相较于母语者，二语学习者的中介语网络呈现出更紧密的连接模式，网络连接更加依赖高度节点，且中介语网络的参数趋近而不超越母语网络。这一现象可能反映了学习者在二语习得过程中采取的策略。

中介语网络具有无标度属性，而无标度网络通常通过“优先连接”（Preferential Attachment）机制来增加新的节点(Barabási and Albert, 1999)，即新节点倾向于与高度节点节点建立连接。“优先连接”会导致网络出现异配性特征。这种网络特性可能反映了学习者在学习过程中的一种自然倾向：将新学习到的词汇与已经熟悉的、使用频率较高的词汇相结合。这可以被理解为认知过程中的一种“锚定”(Tversky and Kahneman, 1974; **?**)。通过将新知识与既有知识关联，学习者能够促进知识的巩固和新知识的内化。然而，这种“锚定”也可能成为障碍，导致学习者过度依赖已有知识，使得其语言水平难以完全达到母语水平。

在二语习得领域，锚定可能与学习者减轻认知负荷的习得模式有关。认知负荷是指在一个特定的作业时间内施加于个体认知系统的心理活动总量(Sweller, 1988)，其假设人的认知结构由工作记忆和长期记忆组成，信息只有先被工作记忆注意和处理后，才能存储在长期记忆中。工作记忆是大脑中负责临时存储和处理信息以执行认知任务的记忆系统，与语言理解、产出和习得等活动密切相关(Bergsleithner, 2010; Wilhelm et al., 2013)。工作记忆理论强调，人们在信息处理过程中可利用的工作记忆容量是有限的；因此，有效的信息组织与处理机制对学习者尤为关键。在学习新词汇时，二语学习者需要依赖工作记忆来暂时存储和处理信息。因此，他们倾向于将新的词汇与已经存在于长期记忆中、容易连接的高频词汇进行搭配编码。这样做可以减轻认识负荷，使得这些词汇更容易在工作记忆中被保持和处理，从而提高学习的效率。一旦这种搭配信息被编码，它就被存储在大脑的长期记忆中。当需要使用长期记忆中的搭配信息时，大脑检索和提取的负担也会更小。即便学习者的语言能力得到提升，这种以节省认知负荷为导向的学习模式仍在其语言习得过程中占据主导地位。例如在C2水平上，学习者所表现出的异配性特征明显超过母语者。这一现象表明，节省记忆负担在二语习得的全过程中都发挥着重要作用。同样地，中介语网络呈现出的层级结构特性，亦反映了学习者减轻认知负荷的学习模式。即使在A1水平，中介语网络已展现出明显的层级结构，该结构提供了一种高效的信息组织模式，使学习者能够通过分层组织词汇的同现搭配，有助于降低语言学习者的认知负担，优化认知资源的使用。

### 4.3 语言错误在网络变异中产生的影响

汉语中介语的语言错误，通常按性质被划分为四大类，分别是遗漏、误加、误代和错序(鲁健骥, 1994)。在写作中，又可以从语言层面上被分成字错误、词错误和句错误(鲁健骥, 1992)。

在中介语网络中，字错误可视为是一种节点内部的错误。这类错误包括错字、别字、漏字、多字等。字错误通常会在网络中引入新的节点。例如，在A2网络中，“时候”这个词，虽然多数学习者正确书写了“时候”，但仍有少数学习者将其错写为“候”。在未考虑书写错误的情况下，“时候”和“候”被视为了不同的节点，从而导致网络节点数量的增加。观察图7（a）也能发现，通过修正这类语言错误，网络的节点数有所减少，纠正后的网络比原网络具有更低的节点数。而中介语网络又是一个复杂自适应系统，语言错误产生的这些新节点，通常又都是低度的，低度节点的涌现，又降低了网络的密度。并且这些低度节点之间相互通达的距离拉长，导致平均路径长度被拉长。同时，低度节点通常都是网络的边缘节点，无法与邻接点形成三元闭包，进而降低了网络的聚集系数。日本汉语学习者在写作中常出现字错误，部分原因在于汉语和日语共用汉字作为文字系统，但两种语言中的汉字并不完全一致。这些易混淆的汉字成为日本汉语学习者常犯字错误的一个原因。例如，汉语中的“年龄”与日语的“年”，汉语的“东京”与日语的“京”都不同，这类错误在网络中增加了大量低度节点，引起网络结构的变异。

词错误不仅影响节点和边的数量，还可能导致节点间的错误连接，包括错词、缺词和多词三种情况。错词是指学习者使用了不恰当的词汇；缺词则是学习者在句子中漏写了必要词；多词则是学习者在书写时多写了不必要的词。错词和多词在网络中都会引入不正确的节点连接，这不仅提高了某些节点的度数，还可能人为地增加了局部聚集系数，因为这些错误连接可能形成了实际上并不存在的三元闭包。此外，错词和多词错误通过引入额外的边，也有可能提升了网络的整体密度。这类错误连接还可能导致两个原本不直接相关的节点之间形成较短的路径，进而减少了网络的平均路径长度。然而，缺词对网络结构的影响则可能与之相反，网络中缺少


(c) 2024 中国中文信息学会计算语言学专业委员会

了原本应当存在的节点，打断了原有的正确连接，因而很有可能导致网络密度的下降，局部聚集系数的降低以及平均路径长度的增加。总而言之，词错误对中介语网络结构的影响是多方面的，它们通过改变节点和边的数量及连接方式，进一步影响网络的统计特性和拓扑结构。这对于理解学习者语言能力的发展和中介语的特性有重要意义。

汉语和日语的语序有很大的不同。日语是黏着语，修饰谓语的词语在句中的位置有较大的自由度(崔立斌, 2001)；而汉语是孤立语，缺乏严格意义上的形态变化，语序则比较固定(叶蜚声、徐通锵, 2015)。因此，受语言迁移的影响，语序错误成为日本汉语中介语中最常见的句错误类型之一(崔立斌, 2001)。在网络分析中，语序错误可被视为关系错误，即边的错误。语序错误对语言网络结构的影响较为复杂，且不易直接从网络的统计特征中捕捉。这类错误一般不会直接增加或减少网络中的节点与边的总数，而是将节点间正确的同现连接改变成错误的同现连接。错误连接的形成以及正确连接的消失，二者共同改变了网络的拓扑结构，这可能会在网络的某些区域造成结构扭曲，从而增加了对语言网络理解的难度。错误的上下文建立连接，可能会导致母语者对中介语内容的理解产生偏差，这也解释了为什么语序错误对作文的可读性影响较大。

值得注意的是，尽管偏误造成了中介语网络的变异，但它们并未改变网络的根本属性，这一点在偏误数量最多的A1网络中尤为明显。A1网络包含的549个偏误，占整个网络词例规模的10.9%。即便如此，A1网络仍然展现出小世界属性、无标度属性以及层级结构特性。这表明，尽管偏误的比例相对较高，它们并未根本影响网络的基础属性，这反映了中介语网络在面对语言偏误时，具有一定的鲁棒性。

## 5 结语

本研究通过深入分析汉语中介语词同现网络，揭示了汉语学习者在汉语二语习得过程中的词汇搭配模式和网络结构变化。本文发现：低水平学习者的中介语网络已经表现出了小世界属性、无标度分布、异配性和层级结构特征，这些特性随着学习者语言水平的提升而逐渐变化。网络分析结果显示：低水平学习者的网络更加紧密并且高度依赖高度节点，反映了学习者减轻认知负荷的习得模式。随着语言水平的提升，学习者开始探索更广泛的词汇组合，网络结构变得更加分散，聚集系数下降，异配性特征减弱，而平均路径长度增加，表明学习者的词汇使用和句法结构变得更加复杂和成熟。此外，语言错误对中介语网络结构的变异也产生了深刻的影响，这些错误不仅增加了网络的节点数，而且影响了网络的密度和聚集系数。

总的来说，汉语中介语词同现网络研究为我们提供了一个全新的视角，可以帮助我理解二语学习者在汉语习得过程中展现的语言习得特征与策略。未来的研究，我们期望通过对比分析不同国别背景下学习者的中介语网络特性及其演变的差异，来揭示多国别背景下学习者的语言发展路径与习得机制，从而为汉语二语习得研究提供更加广泛的实证支持。

## 参考文献

R Albert and A L Barabási. 2002. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47.

A L Barabási and R Albert. 1999. Emergence of scaling in random networks. *science*, 286(5439):509–512.

A L Barabási. 2002. *The new science of networks*. Cambridge MA. Perseus.

JM Bergsleithner. 2010. Working memory capacity and l2 writing performance. *Ciências & Cognição*, 15(2):2–20.

J Borge-Holthoefer and A Arenas. 2010. Semantic networks: Structure and dynamics. *Entropy*, 12(5):1264–1302.

EJ Briscoe. 1998. Language as a complex adaptive system: co-evolution of language and of the language acquisition device. In *Proceedings of eighth computational linguistics in the Netherlands Conference*, volume 4.

L Cameron and D Larsen-Freeman. 2007. Complex systems and applied linguistics. *International journal of applied linguistics*, 17(2):226–240.


(c) 2024 中国中文信息学会计算语言学专业委员会

RFI Cancho, RV Solé, and R Köhler. 2004. Patterns in syntactic dependency networks. *Physical Review E*, 69(5):051915.

J Cong and H Liu. 2014. Approaching human language with complex networks. *Physics of life reviews*, 11(4):598–618.

B Corominas-Murtra, S Valverde, and R Solé. 2009. The ontogeny of scale-free syntax networks: phase transitions in early language acquisition. *Advances in Complex Systems*, 12(03):371–392.

L Danon, AP Ford, T House, et al. 2011. Networks and the epidemiology of infectious disease. *Interdisciplinary perspectives on infectious diseases*.

J Davidsen, H Ebel, and S Bornholdt. 2002. Emergence of a small world from local interactions: Modeling acquaintance networks. *Physical review letters*, 88(12):128701.

NC Ellis and D Larsen-Freeman. 2006. Language emergence: Implications for applied linguistics—introduction to the special issue. *Applied linguistics*, 27(4):558–589.

NC Ellis and D Larsen-Freeman. 2009. Constructing a second language: Analyses and computational simulations of the emergence of linguistic constructions from usage. *Language learning*, 59:90–125.

M Garg and M Kumar. 2018. The structure of word co-occurrence network for microblogs. *Physica A: Statistical Mechanics and its Applications*, 512:698–720.

SA Greenberg. 2009. How citation distortions create unfounded authority: analysis of a citation network. *Bmj*, 339.

Y Hao, X Wang, M Wu, et al. 2021. Syntactic networks of interlanguage across l2 modalities and proficiency levels. *Frontiers in Psychology*, 12:643120.

J Jiang, W Yu, and H Liu. 2019. Does scale-free syntactic network emerge in second language learning? *Frontiers in Psychology*, 10:418102.

M Levine and EH Davidson. 2005. Gene regulatory networks for development. *Proceedings of the National Academy of Sciences*, 102(14):4936–4942.

J Li and J Zhou. 2007. Chinese character structure analysis based on complex networks. *Physica A: Statistical Mechanics and its Applications*, 380:629–638.

W Liang, Y Shi, CK Tse, et al. 2009. Comparison of co-occurrence networks of the chinese and english languages. *Physica A: Statistical Mechanics and Its Applications*, 388(23):4901–4909.

H Liu and J Cong. 2013. Language clustering with word co-occurrence networks based on parallel texts. *Chinese Science Bulletin*, 58:1139–1144.

H Liu. 2008. The complexity of chinese syntactic dependency networks. *Physica A: Statistical Mechanics and Its Applications*, 387:3048–3058.

BH Mayhew and RL Levinger. 1976. Size and the density of interaction in human aggregates. *American Journal of Sociology*, 82(1):86–110.

MEJ Newman. 2002. Assortative mixing in networks. *Physical review letters*, 89(20):208701.

MEJ Newman. 2003a. Mixing patterns in networks. *Physical review E*, 67(2):026126.

MEJ Newman. 2003b. The structure and function of complex networks. *SIAM Review*, 45(2):167–256.

MEJ Newman. 2010. *Networks: an introduction*. Oxford University Press.

Council of Europe. 2018. *Common European Framework of Languages: Learning, Teaching, Assessment*. Cambridge University Press.

GA Pagani and M Aiello. 2013. The power grid as a complex network: a survey. *Physica A: Statistical Mechanics and its Applications*, 392(11):2688–2700.

E Ravasz and AL Barabási. 2003. Hierarchical organization in complex networks. *Physical review E*, 67(2):026112.

CCL 2024

B Richards. 1987. Type/token ratios: What do they really tell us? *Journal of child language*, 14(2):201–209.

J Scott. 2000. *Social Network Analysis: A Handbook*. SAGE Publications.

Larry Selinker. 1972. Interlanguage. *International Review of Applied Linguistics in Language Teaching*, 10(1-4):209–232.

J Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive science*, 12(2):257–285.

A Tversky and D Kahneman. 1974. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157):1124–1131.

S Wasserman and K Faust. 1994. *Social Network Analysis: Methods and Applications*. Cambridge University Press.

DJ Watts and SH Strogatz. 1998. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442.

DJ Watts. 2003. *Six degrees: The science of a connected age*. WW Norton & Company.

O Wilhelm, A Hildebrandt, and K Oberauer. 2013. What is working memory capacity, and how can we measure it? *Frontiers in psychology*, 4:00433.

刘海涛. 2017. 计量语言学导论. 科学出版社.

刘海涛. 2022. 依存关系与语言网络. 科学出版社.

刘涛, 陈忠, and 陈晓荣. 2005. 复杂网络理论及其应用研究概述. 系统工程, (06):1–7.

叶蜚声、徐通锵. 2015. 语言学纲要. 北京大学出版社.

崔立斌. 2001. 日本学生汉语学习的语法错误分析与汉日语言对比. 语言文字应用, (04):20–24.

张宝林、崔希亮. 2022. "全球汉语中介语语料库"的特点与功能. 世界汉语教学, 36(01):90–100.

王士丛. 2021. 基于词同现网络的汉语中介语系统性实证分析. Ph.D. thesis, 华侨大学.

韩笑, 张亮, 张华, et al. 2021. 复杂网络视角的汉语二语口语句法复杂度发展研究. 世界汉语教学, 35(03):377–391.

鲁健骥. 1992. 偏误分析与对外汉语教学. 语言文字应用, (01):69–73.

鲁健骥. 1994. 外国人学汉语的语法偏误分析. 语言教学与研究, (01):49–64.

CCL-2024

# 基于意合图语义理论的结构标注体系与资源建设*

郭梦溪[1]，李梦[1]，荀恩东[2†]，饶高琦[3]，于钟洋[1]
[1]北京语言大学 信息科学学院
[2]北京语言大学 语言资源高精尖中心
[3]北京语言大学 国际中文教育研究院
guo_mengxi@foxmail.com

## 摘要

意合图是一种以事件为中心的多层次语义表示方法，由事件结构与实体结构构成，通过多层次语义体系设计，实现对事件的多层次分析。本文细化并制定了意合图标注规范，采用分层分级的标注策略，在自主研发的在线标注系统中对新闻语料和国际中文教育阅读语料进行了意合图1.0标注工作。通过本次标注，验证了意合图体系的合理性和可标注性，并构建了意合图语义资源库。

**关键词：** 意合图；资源建设；语义表示

# System and Resource Construction Based on the Semantic Theory of Chinese-Parataxis-Graph

**Mengxi Guo[1], Meng Li[1], Endong Xun[2*], Gaoqi Rao[3], Zhongyang Yu[1]**
[1]School of Information Science, Beijing Language and Culture University
[2]Beijing Advanced innovation Center for language Resources,
Beijing Language and Culture University
[3]Research Institute of International Chinese Language Education,
Beijing Language and Culture University
guo_mengxi@foxmail.com

## Abstract

The Chinese Parataxis Graph (CPG) is a multi-level semantic representation method centered around events, consisting of event structures and entity structures. By designing a multi-level semantic system, it enables detailed analysis of events. This paper refines and establishes the annotation standards for the CPG, employing a hierarchical and graded annotation strategy. The CPG 1.0 annotation was carried out on news corpora and international Chinese education reading materials using a self-developed online annotation system. This annotation process validated the rationality and annotability of the CPG system and resulted in the construction of a semantic resource database for the CPG.

**Keywords:** Chinese Parataxis Graph , Resource development , Semantic representation

基金项目：国家自然科学基金“中文意合图的表征与生成方法研究（62076038）；北京语言大学研究生创新基金（中央高校基本科研业务费专项资金）（23YCX136）
† 通讯作者
©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

## 1 引言

在自然语言处理领域，语义分析作为一项重要任务，其核心目标在于从自然语言表达中分析其含义与意图，为计算机理解和处理自然语言提供基础支持。语义表示是将语义分析的结果形式化地进行隐式或显式地表征，隐式表征一般是基于参数的方法进行表征，显式表征一般是基于符号的方法进行表征。目前国内外存在多种符号化的显式语义表示理论，在表示形式和理念上各有不同。基于一种语义表示理论进行语料标注，是对理论科学性的有效验证，也是实现自动分析的数据基础。因此大多语义表示理论提出的同时也构建了相应的语义资源。

意合图是荀恩东近年来提出的一种以事件为中心的多层次语义表示方法，采用单根有向图的形式承载事件、实体、属性及其相互关系，对事件结构与实体结构进行有效表示，实现对事件的多层次分析。本研究基于意合图理论，制定了意合图标注规范，开展了意合图的标注工作，以此对意合图的合理性和可标注性进行验证，并在标注过程中不断打磨与完善意合图理论体系。所构建的语义标注资源可为自然语言处理与语言学研究提供数据支持。

## 2 相关研究

在资源构建方面，意合图在不同发展阶段，积累了具有不同特点的基础资源，为后期基于意合图的语义标注与分析奠定了基础。荀恩东在2019年首届事理图谱研讨会上公开提出意合图理论，随后经历了句法结构树库 (卢露等, 2022)与汉语块依存树库的理论 (钱青青等, 2022)与资源建设 (钱青青等, 2022)，为意合图的逻辑结构知识、论元结构知识和篇章知识打好基础 (钱青青等, 2022)。王诚文 (2021)初次制定了意合图的论元体系，并构建了非主客体论元格标，提供了大规模高质量的动词论元知识。荀恩东 (2023)正式提出了的早期意合图理论架构，并从宏观角度提出了意合图语义分析的资源需求。王贵荣 (2023)面向意合图开展了近一万句论元结构与事件间关系的标注工作，并对事件词与论元的标注情况进行了统计分析，总结了标注不一致的原因，为意合图体系的进一步完善与标注提供支撑。意合图的事件内结构中，除论元结构外，还有情态结构。在自然语言处理领域，相较于论元，对情态信息的研究相对较少，主要集中于情态识别与基于情态的情感分析。邵田 (2023)以大数据为基础构建情态义的状语组块与补语组块语义搭配库，并对其分类体系及语义分布情况进行深入研究，该工作为意合图情态体系的确定与标注提供了理论支撑与数据支持。

意合图经过了逐步探索，在大量实例标注中不断修正认识，完善其体系。在面向意合图构建的论元知识 (王诚文, 2021)与意合图论元标注结果分析 (王贵荣, 2023)的基础上，我们进一步完善了意合图的论元结构；在情态研究 (邵田, 2023)的基础上，我们从计算性与应用性的角度出发，确定了面向意合图的情态体系；考虑到时空信息对事件的重要性，我们对事件的时间信息作进一步细化，构建了意合图的时空结构体系；并且进一步明确了意合图关系事件的定义与作用。此外，除事件结构外，在实体结构方面，我们对实体属性作出通用层面的分类，补充了意合图的实体结构体系。通过上述工作，意合图理论与其语义体系得到进一步完善，目前已初步构建出意合图完整的通用性语义体系。因此，意合图的标注工作从局部标注推进至了整体标注。本文主要从意合图的语义体系内容与当前标注工作的推进两个方面展开。一些最新的相关概念在意合图理论论文 (郭梦溪等, 2024)中已有说明，该文将酌情略写。

## 3 意合图语义标注体系

在计算可行的前提下，基于在通用性下易于扩展的理念，意合图围绕事件表示，将语义分析转化为事件结构与实体结构的分析，设计了多层次通用语义体系，其层次结构如图1所示。本节主要从事件结构与实体结构两部分对意合图语义体系作说明。


(c) 2024 中国中文信息学会计算语言学专业委员会

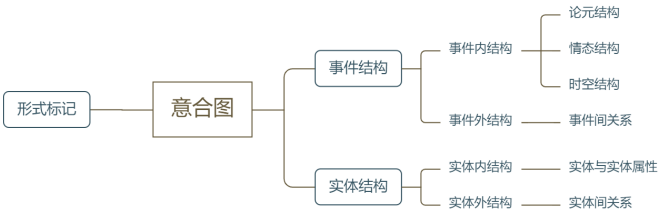



Figure 1: 意合图语义标注体系

### 3.1 事件结构

意合图的事件包括一般事件与关系事件。一般事件是对现实世界或可能世界中事物的动作行为或状态描述。除一般事件外，意合图还将单元的外部关系视作事件，称之为关系事件。 (郭梦溪等, 2024)事件结构分为事件内结构与事件外结构两部分，其中事件词的判断对事件结构的标注，以及标注员对意合图事件的理解尤为关键，因此本节从事件词的判断、事件内结构、事件外结构三个方面展开。

#### 3.1.1 事件词

意合图将事件词作为事件的核心表达，事件词关联起事件内的各个成分，在事件间建立关系时，以事件词代表整个事件。根据事件词是否在句中有对应的语言单元，可以将事件词分为显性事件词和隐性事件词。

显性事件词能够在句子中找到相对应的语言单元，常为谓词，且不受句法位置限制。如“我哭肿了眼”，谓语中心“哭”与补语“肿”均为事件词；“即将远航的列车”与“列车即将远航”中的“远航”均为事件词。

隐性事件词在句子中不能找到对应的语言单元，是人为定义出的概念词语，可分为省略型与关系型两种，省略型事件词是事件词在句中缺省而进行的补全，而关系型事件词是意合图针对关系事件[0]定义出的事件词，是抽象出的关系词语。如下图所示，“目的关系”是代表目的关系事件的事件词，“Or”是代表选择关系事件的事件词。

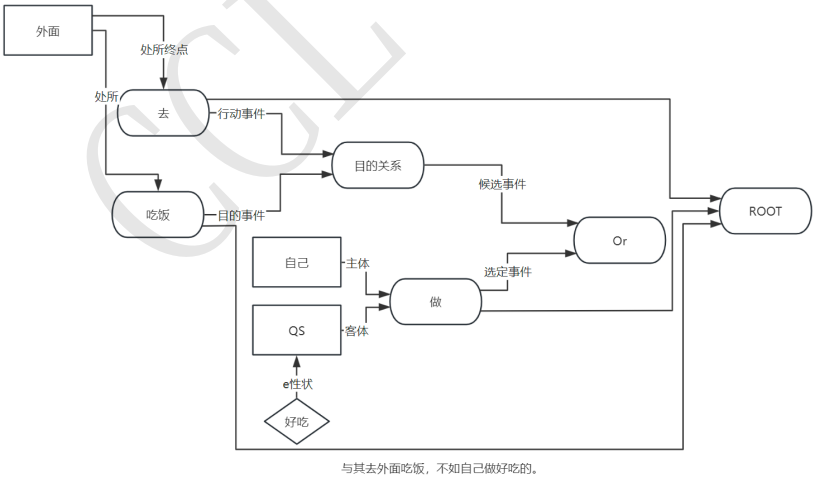



Figure 2: “与其去外面吃饭，不如自己做好吃的”的意合图表示

意合图选择事件词作为事件的核心表达，应选择能够表征事件核心概念的单元，一些不表征事件概念的虚化义谓词不应充当事件词，如形式动词、泛义动词等。其次，意合图的事件词可对应非连续的语言单元，如离合词，在形式化表征语义时，我们将其视为一个非连续性概念单元。对于语义凝固的超词形式的构式等，应视为一个概念单元。

[0]鉴于篇幅问题，本文不再详细解释关系事件，可参考《意合图：中文多层次语义表示方法》


(c) 2024 中国中文信息学会计算语言学专业委员会

### 3.1.2 事件内结构

事件内结构由论元结构、情态结构和时空结构构成，具体如下：

（一）论元结构

论元是事件的构成成分，意合图的论元分为一般论元与关系论元。一般论元是意合图一般事件的构成成分，与学界通常所说的论元角色内涵基本一致，常由实体充当，但非核心事件也可充当论元，此时非核心事件作为整体充当其他事件的构成成分。本次标注的一般论是面向事件词全集定义，不基于特定事件词针对性制定。根据与事件场景的关系，将一般论元角色分为核心论元与边缘论元，核心论元与事件场景紧密相关，边缘论元对事件场景起补充说明作用 (王诚文, 2021)。本研究的论元结构参考学界相关研究，在标注中不断打磨，最终形成包含3个核心论元和13个边缘论元的意合图一般论元体系。其中核心论元采用粗分类，将其聚类为主体、客体与邻体三类，具体情况如表1所示 (王贵荣, 2023)。边缘论元分为13类，具体情况如表2所示。

| 核心论元 | 类型 | 描述 | 示例 |
|---|---|---|---|
| 主体 | 施事 | 自主性动作、行为的发出者 | 小明打了他一下。 |
| | 当事 | 性质、状态或变化性事件的主体 | 宝塔很高。 |
| | 感事 | 非自主的感知性事件的主体 | 我太累了。 |
| | 领事 | 表示领属关系的领有者或整分关系中的整体 | 我有一本书。 |
| | 系事 | 类属关系或比喻关系的主体 | 我是中国人。 |
| 客体 | 受事 | 自发动作行为涉及的已存在事物 | 小明打了他一下。 |
| | 客事 | 非自发动作行为涉及的事件 | 洪水冲毁了房屋。 |
| | 成事 | 由施事的动作、行为造成的结果 | 他写了一本书。 |
| | 属事 | 表示领属关系的所属者，整分关系中的构成部分 | 我有一本书。 |
| | 类事 | 类属关系或比喻关系的客体 | 我是中国人。 |
| | 内容 | 动作、心理涉及但未发生变化的事件或获知信息 | 他知道这个消息。 |
| 领体 | 与事 | 事件中有利害关系的间接客体 | 他送我一束花。 |
| | 同事 | 事件中所伴随的间接客体 | 你跟我去。 |
| | 基准 | 事件中进行比较和测量所参照的客体 | 我比他高。 |
| | 对象 | 事件中动作行为所针对的对象 | 我给她推荐了份工作。 |

Table 1: 核心论元

| 边缘论元 | 描述 | 示例 |
|---|---|---|
| 工具 | 动作、行为所凭借的器具 | 他用毛笔写字。 |
| 材料 | 事件过程中所凭借并消耗的物品 | 我拿面粉做馒头。 |
| 方式 | 事件中主体所采用的方式、方法 | 通过座谈会征询意见。 |
| 依据 | 事件所遵照的根据、条件等 | 依照宪法规定。 |
| 原因 | 引起事件的缘由 | 他以谋杀罪入狱。 |
| 目的 | 事件所要达到的目标 | 为了钱，他不择手段。 |
| 范围 | 事件所作用的领域、范围或覆盖面 | 领导就这个问题发表谈话。 |
| 数量 | 与事件相关的数量、幅度等量值 | 我比他高十公分。 |
| 数量源点 | 与事件相关的数量、幅度等的起始量 | 从50%上调至60%。 |
| 数量终点 | 与事件相关的数量、幅度等的终止量 | 从50%上调至60%。 |
| 状态 | 表示事件发生时主体所处或历经的状态或外部环境情况 | 他努力学习。<br>在逆境中成长。 |
| 状态源点 | 表示事件发生过程中主体的起始状态或外部环境的初始情况 | 从浑浑噩噩到努力学习。<br>从失败走向成功。 |
| 状态终点 | 表示事件发生过程中主体的终止状态或外部环境的最终情况 | 从浑浑噩噩到努力学习。<br>从失败走向成功。 |

Table 2: 边缘论元

第二十三届中国计算语言学大会论文集，第726页-第739页，太原，中国，2024年7月25日至28日。
卷1：主会论文
(c) 2024 中国中文信息学会计算语言学专业委员会

除上述一般论元外，意合图的论元还包括关系论元，关系论元是关系事件的构成成分。意合图将事物间的关系也视为事件，包括实体关系事件与事件关系事件，并对每类关系事件定义了相对应的关系论元。关系论元在事件外结构与实体外结构中进行对应说明。

（二）情态结构

情态是说话人对事件的主观态度，以及事件的客观性抽象特点，是事件的属性信息，依附于事件。在与论元的区分上，情态作为事件的属性信息，强调事件的性质、特点等，而论元是事件的构成成分。从表达情态的词汇载体看，常为副词、助词、形容词等；从句法位置上看，常为状补语；从语义指向上看，大多指向事件词。在意合图中，属性与属性可先建立联系，形成嵌套属性，嵌套属性再依附于事件词，成为事件的属性。如句子“她只会说英语”中，情态信息“只”与“会”发生关系，作为事情词“说”的嵌套属性。意合图目前所定义的情态信息分为8类，分别为：程度、判断、语气、范围、情状、方式、能愿、结果，具体情况见表3所示。

| 情态 | 描述 | 示例 |
|---|---|---|
| 程度 | 事件的动作或状态达到的状况水平 | 这有点好看。我喜欢极了 |
| 判断 | 言者对事件内容的判断结果 | 你的答案不对。 |
| 语气 | 言者对事件的语气表达 | 河水难道倒流吗? |
| 范围 | 事件的涉及范围 | 我只想睡觉。 |
| 情状 | 事件的情形 | 我悄悄地走了出去。 |
| 方式 | 事件的抽象方式特点 | 共同建设美好家园。 |
| 能愿 | 言者的主观意愿或事件的客观可能性等 | 我想去北京。 |
| 结果 | 动作或状态的结果，语义指向事件词 | 我们商量好了。 |

Table 3: 情态

（三）时空结构

时空结构包括与事件相关的时间信息与空间信息。意合图的时间信息目前包括6大类，16小类。具体情况如表4所示；意合图的空间信息目前包括4类，具体情况如表5所示。

| 时间信息 | 类型 | 描述 | 示例 |
|---|---|---|---|
| 时点 | 时间 | 事件发生的时间点或时期 | 他毕业于2023年。 |
| | 时间源点 | 事件开始的时间点或时期 | 他从早上八点学到晚上十点。 |
| | 时间终点 | 事件终止的时间点或时期 | 他从早上八点学到晚上十点。 |
| 时态 | 实现 | 表示事件的发生、出现 | 他去了办公室。 |
| | 进行 | 表示动作的进行或状态的持续 | 我在收拾行李。 |
| | 经历 | 表示曾经发生过某动作，存在某状态 | 我去过北京。 |
| | 将行 | 表示动作行为变化将要发生，或情况状态将要出现 | 我要去北京了。 |
| | 起始 | 表示动作或状态的开始，或进入一个新状态 | 她笑了起来。 |
| | 继续 | 表示已经开始的动作，或状态继续进行或存在 | 我听不下去了。 |
| 时段 | 时段 | 动作或状态持续的时间，或实现后所经历的时间等整段时间 | 我学了一个小时。 |
| 时频 | 时频 | 动作发生的频次或频率 | 他经常迟到。<br>他去过两次。 |

下页继续

Table 4: 时间信息


(c) 2024 中国中文信息学会计算语言学专业委员会

Table 4 – 续表

| 时间信息 | 类型 | 描述 | 示例 |
|---|---|---|---|
| 时序 | 表序 | 表示次序或重复的时间修饰语 | 我回家前，你把屋子打扫了。<br>我先休息一会儿。 |
| | 提前 | 动作发生的时间比预期的时间提前 | 它能够帮你提早发现敌人的踪影。 |
| | 准时 | 动作发生的时间按照预期时间准时发生 | 我按时完成了。 |
| | 延迟 | 动作发生的时间比预期的时间延迟 | 眼下催花技术则更侧重于遮光，让花儿延缓盛开。 |
| 时制 | 时制 | 与时间相关的其他时间修饰语。（不能归属上述语义类别的与时间相关的修饰语） | 早上忽然下起了雨。 |

| 空间信息 | 描述 | 示例 |
|---|---|---|
| 处所 | 事件发生所在或途经的空间位置 | 他坐在椅子上一动不动。<br>他沿着这条路一直走。 |
| 处所源点 | 事件发生的起始空间位置 | 你打哪儿来?<br>我走出教室。 |
| 处所终点 | 事件发生的终止空间位置 | 我走进教室。<br>我回到中国。 |
| 趋向 | 事物随动作而移动的空间位置变化方向 | 他往东走了。<br>他跳下车。 |

Table 5: 空间信息

### 3.1.3 事件外结构

事件外结构即事件与事件间的关系，构成了上文所提到的事件间关系事件。意合图目前定义了8类事件间关系事件，每种事件间关系事件都由其对应的关系论元构成，具体情况如表6所示。

| 事件间关系 | 描述 | 关系论元 | 示例 |
|---|---|---|---|
| 时序关系 | 以时间为坐标轴，表述事件的前后顺序 | 先行事件<br>后继事件<br>伴随事件 | 他吃完饭<先行事件>看电视去了<后继事件>。 |
| 递进关系 | 具有递进关系的两个事件，其中一个事件作为基本事件，另一事件与基本事件相比，在意义上更进一层 | 基本事件<br>递进事件 | 他不但夺得了金牌<基本事件>，还打破了学校的纪录<递进事件>。 |
| 转折关系 | 具有转折关系的两个事件，其中一个事件提出某种事实或情况，另一事件转而述说与前面分句相反或相对的意思 | 让步事件<br>转折事件 | 小明虽然通过了考试<让步事件>，但是他一点也不骄傲<转折事件>。 |

下页继续

Table 6: 关系事件


(c) 2024 中国中文信息学会计算语言学专业委员会

Table 6 – 续表

| 事件间关系 | 描述 | 关系论元 | 示例 |
|---|---|---|---|
| 因果关系 | 具有因果关系的两个事件，其中一个事件表示原因，另一事件表示因此原因而导致的结果 | 原因事件<br>结果事件 | 因为太阳离我们太远了<原因事件>，所以看上去只有盘子那么大<结果事件>。 |
| 条件关系 | 具有条件关系的两个事件，其中一个事件提出影响事件进展的条件，另一事件说明在这种条件下所产生的推论 | 条件事件<br>推论事件 | 如果今天工作不努力<条件事件>，明天就得努力找工作<推论事件>。 |
| 目的关系 | 具有目的关系的两个事件，其中一个事件表示目的，另一个事件表示为达此目的采取的行动 | 目的事件<br>行动事件 | 为了搞好设计<目的事件>，技术人员吊在悬崖上进行工作<行动事件>。 |
| 逻辑关系 | 并列关系 | 并列事件 | 这个西瓜又大<并列事件>又圆<并列事件>。 |
| 选择关系 | 具有选择关系的事件不能并存 | 候选事件<br>选定事件 | 你想逛商场<候选事件>还是看电影<候选事件> |

### 3.2 实体结构

实体结构分为实体内结构与实体外结构。

#### 3.2.1 实体内结构

实体内结构由实体与依附于实体的属性构成。意合图目前将通用层面的实体属性分为10类，具体情况如表7所示。

| 属性 | 描述 | 示例 |
|---|---|---|
| 数量 | 实体的数量、重量、尺寸大小、距离远近等物理属性的度量 | 我买了三斤苹果。<br>现有运输公司32家。<br>很多学生在操场。 |
| 时间 | 实体的存在时间或存在时间的一部分 | 这是唐代的一种风尚。<br>我相信未来的中国一定更好。 |
| 处所 | 实体的所在位置，以及实体的来源 | 公园里的花很漂亮。<br>给我带几瓶绍兴的黄酒。 |
| 参照 | 时间、空间方位、数量的参照对象 | 在两年前的会议上，中国力促亚太自贸区进程启动。<br>我推荐超市旁边的那家店。 |
| 功用 | 实体的用途 | 记者随运动员来到了比赛场地。 |
| 职能 | 实体的职责、职业、产业功能等 | 她是语文老师。<br>现有运输公司32家。 |
| 性状 | 实体内在的性质、质料等，以及实体外在的形态特征等 | 买个塑料凳子。<br>吸引更多优秀的毕业生。<br>会议室摆着一张圆形桌子。 |
| 继续下页 | | |

Table 7: 实体属性


(c) 2024 中国中文信息学会计算语言学专业委员会

Table 7 – 续表

| 属性 | 描述 | 示例 |
|---|---|---|
| 关涉 | 实体所关涉的内容 | 关于七夕的传说。<br>女性文学具有鲜明的女性意识。 |
| 指别 | 具有指别功能的成分 | 我们住在同一个小区。<br>其他人都不去。 |
| 表象 | 当单独看像领属关系，但放入该语境下并非领属关系的，往往是语义上有事件类缺省的特殊情况 | 他的诸葛亮很经典。<br>张三的英语很好。<br>用诗行架起心灵的桥梁。 |

### 3.2.2 实体外结构

实体外结构即实体与实体间的关系，构成了上文所提到的实体间关系事件。在一些情况下，事件可以作为整体与实体发生关系，此时该事件与实体间的关系，相当于实体与实体间的关系。意合图目前定义了5大类实体间关系事件，对每类关系作常见情况列举，细分类为开放类。每种实体间关系事件都由其对应的关系论元构成，具体情况如表8所示。

| 实体间关系 | 描述 | 关系论元 | 示例 |
|---|---|---|---|
| 同指关系 | 同一实体由不同词语指称时，不同词语间具有同指关系 | Entity<br>Event | 李老师<Entity>生病了，他<Entity>女儿很担心。<br>经济上账目不清<Event>的问题<Entity> |
| 领属关系 | 一个实体领有另一实体 | 现领有者<br>原领有者<br>从属者 | 我<原领有者>把书<从属者>送给了他<现领有者>。 |
| 整分关系 | 一个实体是另一个实体的组成部分 | 整体<br>部分 | 我<整体>哭肿了眼睛<部分>。 |
| 并列关系 | 两个或多个实体地位平等 | 并列实体 | 她买了香蕉<并列实体>和葡萄<并列实体>。 |
| 选择关系 | 两个或多个实体间不共存 | 候选实体 | 你喜欢吃香蕉<候选实体>还是葡萄<候选实体> |

Table 8: 实体间关系

### 3.3 形式标记

为了向后期语义分析提供更多的形式化标记，我们对部分句法标记也进行了标注，主要有以下几类：

（一）介词标记

边缘论元常由介词牵引出，因此介词作为格标记，对论元角色的认定起着提示作用，当介引成分为事件时，也对意合图的事件关系认定起重要提示作用，因此介词作为最重要的形式标记之一进行标注。

（二）连接词标记

各种事件关系的连接词对事件关系的认定也同样起重要提示作用，连接词多为连词，也可以为副词等其他具有关联作用的词语。

（三）无实义方位词标记

对于不表示空间位置、时间、数量等的方位词，即无实义的方位词，视为形式标记，进行标注，如“在这个问题上，一定要一刀切”。

（四）话题标记

话题标记是根据对话题的理解和处理而产生的一种语言现象。现代汉语中，句首的“至于”“说到”“再说”等介词或动词后的体词性成分是话题，这些介词或动词性结构经常被称为话题


(c) 2024 中国中文信息学会计算语言学专业委员会

标记。句中语气词之前的体词性成分也是话题，因此这些语气词也可以被认为是话题标记，作为形式标记进行标注。

（五）框式结构标记

标注规范中所允许标注的框式结构以其自身意义凝固为第一原则，如“连……也”“对……而言”；以其引导成分的语义类固定第二原则，如“为了X起见”中的引导成分“X”往往表目的。

### 3.4 特殊现象的处理

#### 3.4.1 省略现象

为保证语义的完整性，意合图允许新增节点进行省略补全，我们定义了事件词省略标记（EW）与实体省略（QS）标记。但由于汉语省略情况复杂，在标注时我们仅补充在该句语义标注中必不可少的语义成分。

事件词补全标记一般用于名词谓语句，除名词谓语句外，只有在句子中某些单元因事件词缺省而导致无法标注进合适的事件结构中时，属于必要性补全情况，如“我一会去超市，你（）呢？”。

实体补全一般有以下几种情况：

（一）“的”字短语

标注中，考虑到语义的明确性，我们将“的”字短语视为中心语省略的定中结构标注，如果在句子中能够找到其所指，补全后与其建立实体间关系以明确指代对象。如句子“红的（）好看”中“的”字短语的标注三元组为：(红,e性状,QS)。

但如果“的”字短语中含有事件表达，则往往不需要进行补全即可标注出其事件结构，该情况实际常出现于“是……的”句的表达中。如句子“饭是中午做的”在意合图中与“中午做饭”的标注一致。

（二）数量短语作为实体属性，实体省略

数量短语后常因表达的经济性省略实体，过往的一些标注常将其视为数量论元与谓词中心进行标注，意合图将其视为实体属性，对缺省的实体进行补充。如句子“我有三个（）”的标注三元组为：[(三个,e数量,QS),(QS,客体,有),(我,主体,有),(有,核心事件词,ROOT)]。

（三）非共识性本体与属性值的搭配

所谓共识性搭配指以实体代实体的通用属性与属性值直接搭配，如“男生很多”“他很高”，以“男生”代“男生的数量”，“他”代“他的个子”，但这种通用属性省略的搭配，已成为一种共识性搭配，一般不会产生语义不明的问题，因此不需要补全。但当句中出现非共识性搭配时，则会造成语义不明或直接推理至共识性搭配的表义，造成语义解析错误，因此应进行省略补全。如句子“她的脸特别红，你（）一点也不红”。

在标注中，标注员需对成分共享与省略补全进行区分，我们从是否出现先行语，以及是否与先行语完全对应的角度对二者进行区别，做不同的标注规范。

当句中无先行语，且符合上述省略补全情况，即可进行省略补全，如“红的没了”。

当承前省略时，多数情况下省略成分与先行语的语义结构是完全相同的，即与先行语完全对应，这时属于成分共享，不需要增添省略标记，如“他用毛巾揩抹了手脸，（他）穿好衣服。”

当省略成分能在句中找到先行语，但与先行语并非完全是同一内容，即与先行语不完全对应，可能是共享部分先行语，或共享全部先行语，但是自身还存在新的属性。在该情况下，如在上述省略补全范围内，则需要增添省略标记，并与先行语建立合适的实体间关系。如句中“四只杯子两只（）碎了”，补全省略实体，并与“四只杯子”建立整分关系。

#### 3.4.2 特殊概念单元

对于汉语中的一些特殊句式，大多仍可以用意合图基本语义体系作出准确表示，但仍有个别句式，我们认为其句式在语义上可拆解为更为清晰，且更有利于下游任务应用的特殊概念单元。

本次标注实践中，我们在比较句中定义了特殊概念单元——比较项目（Comp）。比较句的主要语义是表达参与比较的各方在某一方面比较后的结果 (王贵荣, 2023)，一般由“比较主体、比较对象、比较项目、比较结果”四个部分组成，其中“比较主体”为主体、“比较对象”为邻体、“比较项目”为特殊概念单元标记Comp、“比较结果”为比较句的核心事件词。“比较对


(c) 2024 中国中文信息学会计算语言学专业委员会

象”常由标记词引出，一般为介词，标注为“CompPN”，以此对比较句进行形式标记。上述四个部分在句中可有省略，且比较主体与对象往往共享比较项目。如下例所示：

我的个子比他高。—(我, 主体, 高)、(他, 邻体, 高)、(个子, Comp, 高)、(比, CompPN, 高)

我的个子比他的高。—(我, 主体, 高)、(他, 邻体, 高)、(个子, Comp, 高)、(比, CompPN, 高)

我比他的个子高。—(我, 主体, 高)、(他, 邻体, 高)、(个子, Comp, 高)、(比, CompPN, 高)

我比他高。—(我, 主体, 高)、(他, 邻体, 高)、(比, CompPN, 高)（比较项目省略，属于共识性搭配省略，因此也无需补全）

上述例子中表示同一事件语义，如果根据基本语义体系与标注规范对比较句进行标注，若不考虑比较项目的缺省，其语义表示则不够准确；若考虑缺省情况，则常要面临与先行语不完全对应的省略情况处理，其语义表示不够简洁明了，且同一事件在表示上的一致性难以保证。因此，意合图在语义体系的构建上，在保证其基本原则的前提下，保留其一定的灵活性。

## 4 意合图语义资源建设

### 4.1 语料采样及处理

在语料的采样上，考虑到初期标注语料的规范性与代表性，因此选择了较为规范，且能体现语言生活的国际中文教育阅读语料。其次，考虑到后期能够在同一语料下与其他语义表示方法的标注结果进行有效对比，因此选择了NLP中具有代表性的宾州中文树库CTB（Penn Chinese Treebank）8.0的部分语料。在领域性上，后期将会根据落地场景针对性地选择领域化语料，暂不属本阶段工作范围。在上述考虑下，本次标注最终抽取了1754条来自BCC（Beijing Language and Culture University Corpus Center）(荀恩东等, 2016)国际中文教育语料库的阅读语料，3246条来自宾州中文树库CTB（Penn Chinese Treebank）8.0的新闻语料，共计5000条。

在语料的处理上，在自动抽取后又进行了人工筛选，确保其语法正确、语义自足、语体规范，随后进行分词处理。在分词上，为解决自动分词错误问题，我们在标注平台中设置了分词修正功能（见4.2.2）。

### 4.2 意合图标注实践

#### 4.2.1 标注任务与规范

意合图语义层级分明，框架清晰，因此我们采用了分级分层逐步构建的原则，先行建立意合图框架，再逐一根据需求细化各部分内部结构。本次标注工作旨在建立意合图框架，标注内容包括细粒度的论元结构、粗粒度的情态结构、细粒度的事件间关系事件、粗粒度的实体内结构，而实体间关系事件仅标注了同指关系事件，领属关系事件与整分关系事件同实体内结构做一致标注。

分层分级展开标注能够减轻标注员的标注负担，合理控制标注员记忆内容以符合科学的标注心理，以此提高标注效率与质量。但我们在制定规范时，均从细粒度出发，标注员在前期培训中，也按照细粒度语义标签进行标注，以此加深标注员对每个语义标签的理解，避免粗粒度标注任务使标注员对语义体系的理解不够深入。因此，标注员即使在该阶段进行粗粒度标注，但其查阅规范时可看到详细的细粒度语义情况。

标注规范是资源建设的关键环节，对标注的准确性和一致性至关重要。本次任务的标注规范基于意合图语义体系展开，并根据标注情况动态修改或增加细则，不断完善规范内容。截至2024年4月，意合图标注规范共有七章，除上述基本语义体系内容外，还包括复杂情况的标注示范、标注过程中的常见问题汇总与阶段性质检反馈文档，整体内容近四万字，采用在线文档的方式将更新内容实时同步给标注员。

#### 4.2.2 标注工具与流程

为提升标注效率和质量，我们自主开发了在线标注系统（图3）。通过边缘计算的方式，将用于检查标注结果的构图计算环节移动到终端设备进行，以提高响应速度并减少服务器的压力，从而支持系统处理大规模数据。该在线标注系统既可导入分词后的生语料，也可以导入标注过的语料进行二次标注。为了减少非认知性的不一致标注结果，我们在网页中加入提示语，并且在操作上自动排查一些常见错误，例如每条语料中至少有一个核心事件词，如检测到未标注核心事件词，在点击一下句时会自动提示返回上句标注核心事件词；对于具有对应关系的情


(c) 2024 中国中文信息学会计算语言学专业委员会

况，设置其对应内容，也进一步提升了标注效率。标注是以三元组标注的方式进行，但三元组集合是面向机器的表示方法，对于人并不友好，标注员不易通过三元组集合构建出整个标注对象内的语义关系，因此我们加入了根据标注三元组动态生图的功能，以更友好的方式呈现给标注员现标注情况，以便于快速排查错误。




Figure 3: 标注平台页面示例

本次标注由8位具有语言学专业背景的研究生在标注规范的指导下进行。我们本着“少量精标”的理念，采用两两对比的形式进行标注，每次任务由两位标注员独立标注同一份语料，标注完成后返回对比结果，双方对不一致结果进行讨论，确定唯一标注结果。如双方不能达成一致，则由专家（管理者）介入，确定标注结果。最终所提交的一致标注结果由管理者再次进行质检校对后入库。管理者除查看双方不一致标注结果与质检外，还阶段性参与标注，及时发现标注规范的不足与标注员存在的问题。因此，在标注平台的辅助下，同一份语料实际经过3轮、4次标注，且管理者通过多种方式深入标注工作，一定程度上保证了其标注质量。

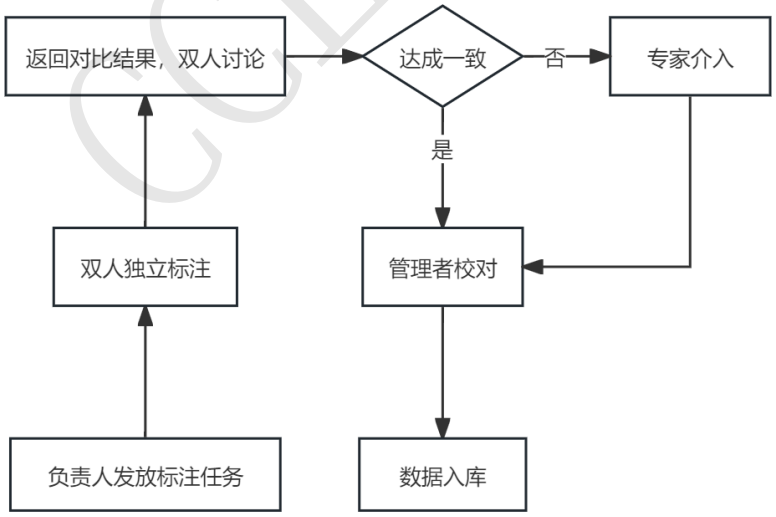



Figure 4: 标注流程图

### 4.2.3 标注数据与结果

标注数据共包含5000条语料，总计11.5943万字，平均每条语料有13.89个词，23.19个字。

国际中文教育领域的阅读语料平均每条12.82个词，19.22个字；来自宾州中文树库的部分新闻语料平均每条14.46个词，25.33个字。我们对标注结果进行进一步统计，其分布情况如表9、10所示。

| | 阅读文本 | 阅读文本句均 | 新闻语料 | 新闻语料句均 | 总体 | 总体句均 |
|---|---|---|---|---|---|---|
| 主体 | 3244 | 1.85 | 5973 | 1.84 | 9217 | 1.84 |
| 客体 | 2301 | 1.31 | 5184 | 1.60 | 7485 | 1.50 |
| 邻体 | 261 | 0.15 | 543 | 0.17 | 804 | 0.16 |
| 工具 | 47 | 0.03 | 39 | 0.01 | 86 | 0.02 |
| 材料 | 3 | 0.00 | 9 | 0.00 | 12 | 0.00 |
| 方式 | 74 | 0.04 | 191 | 0.06 | 265 | 0.05 |
| 依据 | 23 | 0.01 | 186 | 0.06 | 209 | 0.04 |
| 原因 | 43 | 0.02 | 53 | 0.02 | 96 | 0.02 |
| 目的 | 0 | 0.00 | 14 | 0.00 | 14 | 0.00 |
| 范围 | 77 | 0.04 | 390 | 0.12 | 467 | 0.09 |
| 数量 | 18 | 0.01 | 191 | 0.06 | 209 | 0.04 |
| 数量源点 | 1 | 0.00 | 20 | 0.01 | 21 | 0.00 |
| 数量终点 | 2 | 0.00 | 116 | 0.04 | 118 | 0.02 |
| 状态 | 87 | 0.05 | 144 | 0.04 | 231 | 0.05 |
| 状态源点 | 5 | 0.00 | 14 | 0.00 | 19 | 0.00 |
| 状态终点 | 33 | 0.02 | 45 | 0.01 | 78 | 0.02 |
| 情态 | 2349 | 1.34 | 2054 | 0.63 | 4403 | 0.88 |
| 实体结构 | 3139 | 1.79 | 9909 | 3.05 | 13048 | 2.61 |
| 核心事件词 | 2837 | 1.62 | 4357 | 1.34 | 7194 | 1.44 |

Table 9: 标注情况统计

| | 阅读文本 | 新闻语料 | 总计 |
|---|---|---|---|
| 时序关系 | 262 | 275 | 537 |
| 递进关系 | 30 | 59 | 89 |
| 转折关系 | 52 | 51 | 103 |
| 因果关系 | 237 | 183 | 420 |
| 条件关系 | 135 | 81 | 216 |
| 目的关系 | 144 | 202 | 346 |
| 并列关系 | 337 | 1172 | 1509 |
| 选择关系 | 37 | 42 | 79 |
| 同指关系 | 224 | 588 | 812 |

Table 10: 外部关系标注情况统计

由于意合图是以事件为中心的语义表征图，我们对标注数据中所含事件词情况进行了统计。统计得到，该标注数据中有12812个不同的事件词（不包含隐性事件词）。在国际中文教育领域的阅读语料中出现次数前五的事件词为“是”“有”“说”“到”“去”；在来自宾州中文树库的部分新闻语料中出现次数前五的事件词为“是”“说”“发展”“有”“投资”；总标注语料中出现次数前五的事件词为“是”“有”“说”“发展”“投资”。

## 5 结语

本文描述了意合图语义体系的具体内容，并介绍了基于意合图语义体系在展开的标注工作，对意合图语义体系进行打磨，验证了其合理性与可标注性，并构建了一批语义标注资源，为后期意合图自动生成提供数据基础。该资源也开源投入评测活动，供学界同仁使用。


(c) 2024 中国中文信息学会计算语言学专业委员会

在接下来的工作中，我们将进一步审视标注结果，打磨意合图理论，为意合图的应用探索可行途径，并基于后续应用需求制定2.0计划。

## 致谢

意合图1.0标注工作由北京语言大学李梦、何晴、胡星雨、王静怡、吴晓靖、张可芯、周书帆、朱奕瑾（按姓氏排）八位研究生完成，在此表示感谢。感谢匿名审稿人对论文提出宝贵的修改意见。

## 参考文献

Bin Li, Yuan Wen, Lijun Bu, Weiguang Qu. 2016. *Annotating the Little Prince with Chinese AMRs.* In Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016), pages 7–15, Berlin, Germany. Association for Computational Linguistics.

Jingyi Wang, Endong Xun. 2023. Construction and Statistical Analysis of Discourse Cohesive Components in Modern Chinese. International Journal of Asian Language Processing, 33(04).

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. *The Proposition Bank: An Annotated Corpus of Semantic Roles.* Computational Linguistics, 31(1):71–106.

Yihuan Liu, Bin Li, Peiyi Yan, Li Song, Weiguang Qu. 2016. *Ellipsis Annotation and Statistics in Chinese based on Chinese AMR.* International Journal of Knowledge and Language Processing.

陈立民. 2002. 汉语的时态和时态成分. 语言研究.

龚千炎. 2004. 现代汉语的时间系统. 语言文字应用研究论文集（Ⅱ）.

龚千炎. 2012. 汉语的时相、时制、时态. 商务印书馆, 北京.

郭梦溪, 荀恩东, 李梦, 饶高琦. 2024. 意合图：中文多层次语义表示方法. 第二十三届中国计算语言学大会.

黄彤, 李斌, 闫培艺, 戴玉玲, 曲维光. 2020. 基于抽象语义表示的汉语构式标注与分析. 中文信息学报.

卢露, 矫红岩, 李梦, 荀恩东. 2022. . 基于篇章的汉语句法结构树库. 自动化学报,48(12): 2911–2921.

刘亚慧, 杨浩苹, 李正华. 2020. 一种轻量级的汉语语义角色标注规范. 中文信息学报.

钱青青, 王诚文, 王贵荣, 饶高琦, 荀恩东. 2022. 基于组块分析的汉语块依存语法. 中文信息学报, 36(08): 20–28.

钱青青, 王诚文, 荀恩东, 王贵荣, 饶高琦. 2022. 汉语块依存语法与树库构建. 中文信息学报, 36(07): 50–58.

邵田. 2023. 基于大数据的汉语情态义简单状补组块研究. 北京语言大学博士论文.

邵田, 翟世权, 饶高琦, 荀恩东. 2023. 基于结构树库的状位动词语义分类及搭配库构建. 中文信息学报,37(06):44-51+66.

邵艳秋, 邱立坤, 梁春霞, 毛宁. 2011. 中文语义依存树库构建及自动分析技术. 中国中文信息学会.中国计算语言学研究前沿进展（2009-2011）.

宋衡, 曹存根, 王亚, 王石. 2023. 一种改进的汉语语义角色分类体系与标注实践. 中文信息学报.

田思雨, 邵田, 荀恩东, 饶高琦. 2023. 基于结构树库的补语位形容词语义分析及搭配构建. 第二十二届中国计算语言学大会论文集,第420页-第432页,哈尔滨.

王诚文. 2021. 面向意合图的汉语动词论知识构建研究. 北京语言大学博士论文.

王诚文, 钱青青, 荀恩东, 邢丹, 李梦, 饶高琦. 2020. 三元搭配视角下的汉语动词语义角色知识库构建. 中文信息学报.

王贵荣. 2023. 意合图事件结构标注及分析研究. 北京语言大学博士论文.

文贞惠. 1999. *"N_1(的)N_2"*偏正结构中*N_1*与*N_2*之间语义关系的鉴定. 语文研究.


(c) 2024 中国中文信息学会计算语言学专业委员会

荀恩东. 2023. 自然语言结构计算：意合图理论与技术. 人民邮电出版社.

荀恩东. 2023. 自然语言结构计算：*BCC*语料库. 人民邮电出版社, 北京.

荀恩东, 饶高琦, 肖晓悦, 臧娇娇. 2016. 大数据背景下BCC语料库的研制. 语料库语言学.

萧国政. 2020. 语法事件与语义事件——面向人工智能的语言研究. 长江学术.

朱德熙. 1982. 语法讲义. 商务印书馆, 北京.

张牧宇, 秦兵, 刘挺. 2014. 中文篇章级句间语义关系体系及标注. 中文信息学报.

周国光. 2002. 现代汉语的语义属性系统. 世界汉语教学.

CCL 2024

# 意合图：中文多层次语义表示方法*

**郭梦溪**[1]，**荀恩东**[2†]，**李梦**[1]，**饶高琦**[3]
[1]北京语言大学 信息科学学院
[2]北京语言大学 语言资源高精尖中心
[3]北京语言大学 国际中文教育研究院
guo_mengxi@foxmail.com

## 摘要

基于参数的语义表示虽取得成就，但符号化的语义表示仍具有不可忽视的意义。我们在语义学基础上，充分考虑符号化语义表示在NLP领域落地中的需求，提出了一种兼具通用性与扩展性的多层次语义表示方法——意合图。意合图以事件为核心，由事件结构与实体结构构成，通过多层次语义体系设计，提升与场景结合的能力，并力求对不同层级的语言单元作一贯式表示。在资源建设和相关分析实验中取得良好效果。本文将重点介绍意合图设计理念与多层次语义体系。

**关键词：** 意合图；语义表示；事件结构；实体结构

# Parataxis Graph: Multi-level Semantic Representation Method for Chinese

**Mengxi Guo**[1], **Endong Xun**[2*], **Meng Li**[1], **Gaoqi Rao**[3]
[1]School of Information Science, Beijing Language and Culture University
[2]Beijing Advanced innovation Center for language Resources,
Beijing Language and Culture University
[3]Research Institute of International Chinese Language Education,
Beijing Language and Culture University
guo_mengxi@foxmail.com

## Abstract

While achievements have been made in parameter-based semantic representation, the significance of symbolic semantic representation cannot be overlooked. Based on semantic theory, this paper thoroughly addresses the requirements for symbolic semantic representation in the application within the NLP field, introducing a universal and extensible multi-level semantic representation method named the Chinese Parataxis Graph. Centering on events, the Chinese Parataxis Graph is structured around event and entity frameworks, enhancing the capability to integrate with scenarios through a multi-level semantic system design and consistently representing language units at various levels. The approach has shown promising results in resource development and related analytical experiments. This paper will primarily discuss the design philosophy and the multi-level semantic system of the Chinese Parataxis Graph.

**Keywords:** Chinese Parataxis Graph , Semantic representation , Event structure , Entity structure

基金项目：国家自然科学基金“中文意合图的表征与生成方法研究（62076038）；中央高校基本科研业务费（北京语言大学梧桐创新平台，21PT04）
† 通讯作者
©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

## 1 引言

从自然语言处理的方法看，语义表示可分为符号化的语义表示与参数化的语义表示。符号化的语义表示利用离散的符号来表征自然语言的语义结构，强调语义结构的逻辑和形式化的显式表示；而参数化的语义表示则通常由向量的分布特征来表示，通过神经网络技术自动学习语言单元在连续向量空间中的分布关系来捕捉语义信息，往往被视为隐式表示。这两种表示方法在不同的任务和场景中都有其各自的适用性。

尽管参数化的语义表示方法已取得显著成就，但符号化的语义表示方法仍然具有其独特的优势和适用场景。荀恩东 (2023)正式提出中文意合图的早期理论架构，以事件为中心，通过图结构的形式实现对事件的多层次表示。在意合图中，事件和实体被动态地编码为节点，通过有向边的连接来表示事件、实体、属性的语义关系，构建多层次的语义结构图。在意合图理论的基础上，我们通过标注近10000句论元结构 (王贵荣, 2023)、5000句意合图框架 (郭梦溪等, 2024)对意合图理论进行大规模验证与修正，在工程实践中我们优化了意合图早期理论架构，完善了意合图通用语义体系。本文将重点介绍意合图的设计理念，以及多层次的体系内容。

## 2 相关研究

本文所研究的意合图是一种符号化的语义表示。截至目前，学界已积累了大量从不同视角设计符号化的语义表示方法的研究，国内外广泛流行的语义表示有多种形式。Fillmore (1977)提出框架语义理论，通过“框架”对语言单元的概念结构进行描述。如动词“取”的语义框架由接受者、转移体、方式、处所等框架元素构成。加州大学构建英语的框架语义网（FrameNet） (Collin F et al., 1998)；国内山西大学借鉴FrameNet，构建了汉语框架语义网（Chinese FrameNet） (刘开瑛, 由丽萍, 2006; 郝晓燕等, 2007)，截至目前共为21283个词汇定义了1327个框架[0]。然而，尽管基于框架的语义分析能提供更精准的语义信息，但建立与维护需要消耗大量的人力和时间，且一旦定义后其泛化能力有限，缺乏灵活性。语义角色与框架元素类似，但框架元素是对具有相同框架的词汇而针对性定义的，语义角色是面向普遍词汇的，相比之下，语义角色标注（Semantic Role Labeling, SRL） (Martha Palmer et al., 2005)更具直接应用性。但语义角色标注仅关注谓词与其他成分的关系，语义覆盖度有限，属于浅层语义分析，且常依赖于句法结构树，语义表示上受句法限制。随着自然语言处理的不断发展，对语义分析的广度和深度日益提升，又陆续出现语义依存分析、抽象语义表示、通用认知概念标注等深层语义分析。语义依存分析（Semantic Dependency Parsing, SDP） (Robinson, 2005)对所有词对之间可能存在的语义关系进行表示，随着由语义依存树转换为语义依存图，语义依存分析打破树结构的限制，使其能够覆盖更多的特殊语言现象。李明琴等 (2003)最早进行汉语的语义依存结构标注。但语义依存对自然语言的语义表示力度仍存在局限，对诸如省略、共指等重要的语言现象无法进行有效表示，并缺乏对不同句式的一致性表示。抽象语义表示（Abstract Meaning Representation，AMR）则采用图结构形式对句中概念与概念之间的关系进行表示，且忽视语义较虚的部分，允许新增节点，使得所能表示的语言现象大大增加 (Laura Banarescu et al., 2013)。南京师范大学将其引入汉语 (Bin Li et al., 2016)，并融入概念对齐信息 (Bin Li et al., 2019)，大大提高了汉语语义表示的深度与广度。但由于抽象语义表示将词语抽象为概念进行表示，存在概念对齐问题，且语义体系复杂，对众多非核心语义关系没有进行层次化分类。通用认知概念标注（Universal Conceptual Cognitive Annotation, UCCA） (Omri Abend et al., 2013)是一种跨语言适用的多层次语义表示，多层次的语义体系使其具有更强的扩展性，但由于其跨语言跨领域适用特点，使其语义表示粒度较粗。

在探讨语义表示的同时，我们不可避免地会涉及到知识表示。目前的知识表示可分为多种类型，包括知识图谱、事理图谱等。知识图谱主要用于表示实体及实体间的关系，而事理图谱是对事件之间的演化规律与模式的常识知识表示 (刘挺, 2017)。尽管传统的语义表示与知识表示的侧重点有所不同，前者侧重于关注语言单元及单元关系，后者侧重于构建一般性知识的结构化表示，但两者在本质上均可视为对概念单元及相互关系的表示，这意味着二者可以在表示逻辑上接近，在表示方法上相互贯通，从而实现有效融合，以提高语义分析与场景结合的能力。

综上所述，不同的语义表示各有优势，又各有局限性。高精度的语义表示能够更精准地描述语义，但往往会导致资源构建成本较大；其次，理想的语义表示应当能够广泛覆盖各种语言

[0]数据参考来源：http://www.cfn-lab.com/#/resourceFrame


(c) 2024 中国中文信息学会计算语言学专业委员会

现象，但也常会使语义体系复杂化；此外，一个高通用性的语义体系可以广泛应用于各种任务，但同时也应保持开放扩展性，以适应不同的应用需求。过往汉语的语义表示通常源于英语的表示方法，在一定程度上受形式性语言的表示方案思路限制。通过审视各种语义表示的优势和局限，以及知识表示的特点，我们针对汉语提出一种具有通用性与扩展性的多层次语义表示方法——意合图。

## 3 意合图的理念

各类语义表示方法无不根植于其语言观和应用观，并在计算可行性的约束下形成具体的表示方案。鉴于高精度的语义表示虽能更精确地描述语义却带来高成本的资源构建问题、理想的语义表示覆盖广泛语言现象而易导致体系过于复杂化的挑战、以及一个具有通用性的语义体系需同时保持可扩展性以适应不同应用的需求，我们将句子信息分为事件信息和实体信息，并采用单根有向图的形式承载事件、实体、属性及其关系，通过多层次语义设计以及与知识表示对接，使其在表示精度与构建成本，语义覆盖度与体系复杂度，通用性与扩展性上取得良好的平衡，进而从应用性角度提升了语义分析与场景结合的能力，我们将其定义为意合图。

### 3.1 语言观

语言观也即整个意合图理论所基于的对语言事实的基本假设。意合图的语言观源于以事件来理解语言符号从而构建对现实、认知、语言三个平面的一致性分析通道，形成了意合图以事件为中心的表示方法，并在表示方法与语义体系的构建上符合语言符号在不同视角下所具有的层次性特点。

#### 3.1.1 人类语言的认知特点

人类在对外部世界种种现象的感知体验中逐步形成范畴、概念和思维，抽象出认知模型，建立认知结构，获得意义，通过语言形式将意义固定下来 (王寅, 2021)。语义形成的过程即可视为范畴化、概念化的认知过程。人类对外部世界感知后进行信息加工的本质是对事件内容的加工，客观存在的事件经感知后形成事件概念，并通过语言进行表达，形成所谓的语言事件 (Rosen, 2003)。因此，“事件”有客观存在的事件、存在于个体心智中的事件以及经语言表征出的事件三种形态，事件结构便可成为对现实、认知、语言三个平面的一致性分析对象，为深度语义分析提供有效视角。因此，意合图采用以事件为中心的语义表示方法符合人类对语言的认知特点。

意合图选择事件词作为事件的核心表达。所选择的事件词应是能够表征该事件的核心概念单元，与语言符号单元不存在绝对的对应关系。例如，一些虚化的谓词不表征事件概念，如形式动词；概念单元所对应的语言单元可以是非连续的，如汉语中的离合词，英语中的分隔动词短语；被恰当选择表示的最小概念单元不一定是词，如一些以短语形式表达的惯用语、构式等所表征的意义不宜再进行拆分。除此之外，同一事件可以用不同句法形式进行表征，在语义表示时应体现出其语义的一致性。

意合图基于这种认识，在语义表示的设计上以事件为中心，突破句法形式，忽视虚化成分，允许位于任何句法位置的语言单元充当事件词，允许多个连续或非连续语言单元充当事件词，对于隐性语义甚至可定义未在句中出现的概念词语作为事件的核心表达，进一步深化了汉语语义表示的深度。

#### 3.1.2 语言符号层次性特点

语言有不同层级的语言单元，如词、短语、句子、段落、篇章等。意合图通过定义与分析关系事件（见5.1），为理解和表示语言的复杂层次提供了新的视角，不仅能更准确地表示语义信息，还实现了对不同层级的语言单元进行一贯式表示。

以句子“我没有出去玩”为例，传统语义表示方法往往将“没有”就近与“出去”直接关联（如图1[1]），这种处理未对否定词的真正作用范围加以考虑，可能导致语义理解上的偏差。而意合图在无上下文指明“没有”的确切语义指向时，将“没有”与“出去玩”这一目的关系事件建立联系（见图2），更符合“我|没有||出去|||玩”的层次格局。

[1]该语义依存图由LTP平台 (Wanxiang Che et al., 2021)自动生成


(c) 2024 中国中文信息学会计算语言学专业委员会




Figure 1: “我没有出去玩”的语义依存图

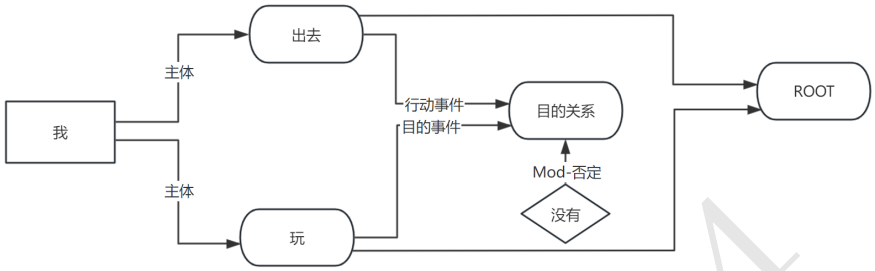



Figure 2: “我没有出去玩”的意合图

对于实体结构也同样如此，以“我和他的关系”及“我和他的书[2]”为例，二者在句法格式上类似，但意合图中所形成的抽象语义表示形式却不同。前者的层次为“我和他（的）|关系”，而后者是由“我的书”与“他的书”构成并列的逻辑关系，二者对应的意合图如图3所示。

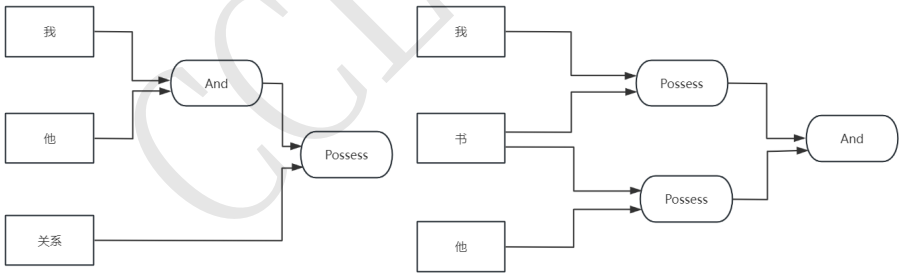



Figure 3: “我和他的关系”与“我和他的书”的意合图对比

除此之外，意合图同样通过关系事件，将不同层级的语言单元作一贯式统一表示，如“小明去商店给我买了一根雪糕，但是天太热化了。他很愧疚……”的意合图如图4所示，通过关系事件对大大小小的语言单元层层组装，实现一贯式和层次化表示。

[2]该处特指在类似“这些是我和他的书”的语境下，即排除“我（和）—他的书”的歧义

第二十三届中国计算语言学大会论文集，第740页-第749页，太原，中国，2024年7月25日至28日。
卷1：主会论文
(c) 2024 中国中文信息学会计算语言学专业委员会

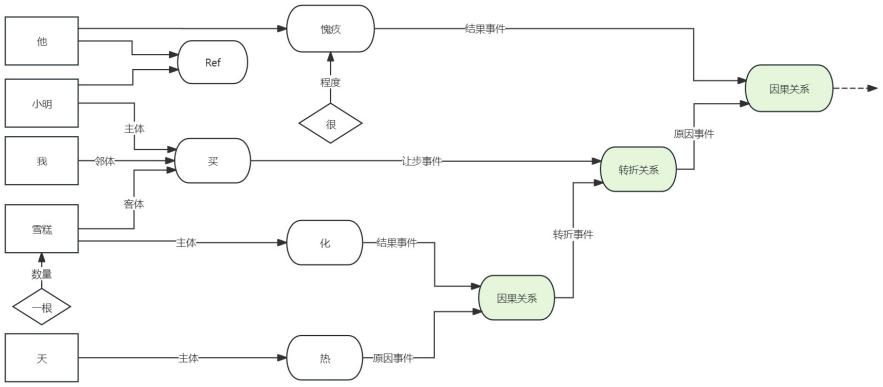



Figure 4: “小明给我买了一根雪糕，但是天太热化了。他很愧疚……”的意合图

语义信息也同样具有不同层级。意合图对语义信息的层次化处理一方面体现在区分事件结构与实体结构，以及这两种结构内部更深层次的划分；另一方面体现在与命题和超命题的层次对应，语义角色标注等浅层语义分析往往仅对命题结构作分析，而语义依存分析、抽象语义表示等深层语义分析虽然已将语义表示范围扩展到了超命题结构，但往往未明确区分命题结构与超命题结构，而意合图将超命题结构归于情态信息与时体信息中，实现了对命题与超命题结构的层次区分。

### 3.2 应用观

应用观也即整个意合图理论希望服务下游NLP任务的形式。意合图的应用观立足于对语义的认识，在应用可行性的前提下采用更为灵活的策略扩大语义覆盖度，致力于建立一个具有通用性与可扩展性的语义体系，在自底向上的知识构建理念下增强与知识表示的兼容性。

#### 3.2.1 语义的覆盖度

意合图不仅对通过语言符号显化的事件结构进行表示，还考虑隐性的语义信息。当人们使用语言对事件进行表征时，通常只会突显事件的一部分。例如句子“他去了”，在语境以及考虑表达突出点的条件下，该事件的起点、目的地、方式等事件要素均未以语言符号的形式进行表征；除事件要素外，一些事物关系也常不以语言符号的形式直接描述，如句子“他哭肿了眼睛”，“他”和“眼睛”的领属关系，“哭”和“肿”的因果关系并没有以语言符号形式进行描述。尽管语言符号形式不会表征出事件的全部内容，但听者在对语言信息进行解码时会激活对该事件认知的必要内容。这就要求在自然语言处理中，语义表示不仅要捕捉通过语言符号直接表达的内容，也需要深入挖掘和形式化表征那些隐性的、未被直接表述的语义信息。

隐性语义信息的补全高度依赖于具体场景与常识知识，仅依靠单一的语义表示很难完成。因此意合图在设计时并没有追求将隐性语义信息完全显式化，而是采用了更灵活的策略。基于出现的语言符号单元，一方面通过定义关系事件，显式化单元之间的隐性关系，如实体间关系和事件间的关系；另一方面，如果因省略等情况而导致某些语言单元无法与其他单元建立合适的语义关系[3]，如“我一会儿去超市，你（∅）呢？”，意合图则通过新增节点来进行补全。对于其他隐性语义信息，意合图可进一步通过外接知识图谱等知识源来进行补充。对此，意合图在设计之初就充分参考了知识图谱和事理图谱等表示方法，使其能够更好地与之兼容。

#### 3.2.2 通用性与扩展性

意合图的通用性体现在表示跨领域甚至跨语言的基本语义关系。这种通用性来源于其设计初衷，即从一个全局视角出发，定义那些在多种领域和场景下普遍存在的语义关系，例如，实体之间的领属关系，在不同领域中可能表现为更具体，且具有显著特点和高应用价值的形式，如所有权关系、物种与其类群等，但这些都可以归纳为通用视角下的领属关系。这样的设计不仅使得意合图能够适用于广泛的应用场景，还为不同领域之间的知识迁移和整合提供了一个共同的语义基础。

[3]关于意合图对省略问题的处理，于意合图资源建设论文中详述。


(c) 2024 中国中文信息学会计算语言学专业委员会

意合图的扩展性则源于其层次分明的结构设计，这种多层次设计使得意合图能够灵活应对各种具体的语义表示需求。通过在必要时调整层次结构或增加新的层次，实现对不同层次和粒度的语义表示需求的灵活适应与扩展。

意合图通过其通用性和扩展性的结合，提供了一个灵活的语义表示框架。通用性使得它能够作为基础工具跨领域应用，而扩展性则保证了其在面对特定需求和不断发展的语义分析需求时的适应性和进化能力。这两大特性共同使得意合图有助于连接不同知识领域、促进跨领域知识迁移和整合。

### 3.2.3 实例到抽象转换

意合图的设计理念强调“先有实例，再做抽象”。这种理念本质上是一种自底向上的知识构建策略，与之相对的是自顶向下的概念先行策略。构建概念体系通常需要大量资源，并且一旦建立，其缺乏灵活性，在适应新信息时成为明显短板。通过大量实例抽象出概念，可以大大减少预先劳动；且在实例驱动的模式下，随着新实例的加入，可动态调整知识体系。此外，意合图在设计上借鉴了知识图谱、事理图谱等知识表示方法，增强了与现有知识体系的兼容性，使得意合图不仅可以从现有的知识图谱和事理图谱中快速提取和整合知识作为补充，还能将从文本实例中新抽象的概念信息有效地整合进这些图谱框架中，实现从实例到概念，再到知识图谱的流畅转换。

## 4 意合图的表示方案

意合图面向人以更友好的图形式表示语义信息，同时也可转换为三元组集合面向机器输入。

### 4.1 单根有向图

意合图是以事件为中心的语义表征图，为单根有向图。图中的节点表示单元，对应承载实体、事件、属性的语言单元，其中我们以事件词作为事件的核心表达，因此节点的语言单元体现为实体词、事件词、属性词。边为有向边，表示单元间的语义关系，对应实体与事件的关系、属性与实体的关系、属性与事件的关系、实体与实体的关系、事件与事件的关系，以及属性与属性的关系。并且设置了“ROOT”节点作为根节点，由核心事件词指向“ROOT”根节点。意合图的抽象表示形式如图6所示。

意合图主要由事件结构和实体结构构成。事件结构（见图5.a）包含内外两部分，在意合图中事件词的上方、下方、左方表示的是事件内结构，是事件词三个方向的输入，为事件的内部构成成分与依附于事件的属性。左方输入是核心论元，上方输入是边缘论元，下方输入是情态信息和时空信息。在意合图中事件词的右方表示的是事件外结构，是事件词的输出。当事件词为核心事件词时，右侧与“ROOT”根节点连接。实体结构（见图5.b）也包含内外两部分，在意合图中实体只有下方输入和右方输出。下方输入实体的属性，是实体内结构；右方为实体词的输出，是实体外结构。




(a) 意合图事件结构的抽象表示　　(b) 意合图实体结构的抽象表示

Figure 5: 意合图事件结构与实体结构的抽象表示


(c) 2024 中国中文信息学会计算语言学专业委员会

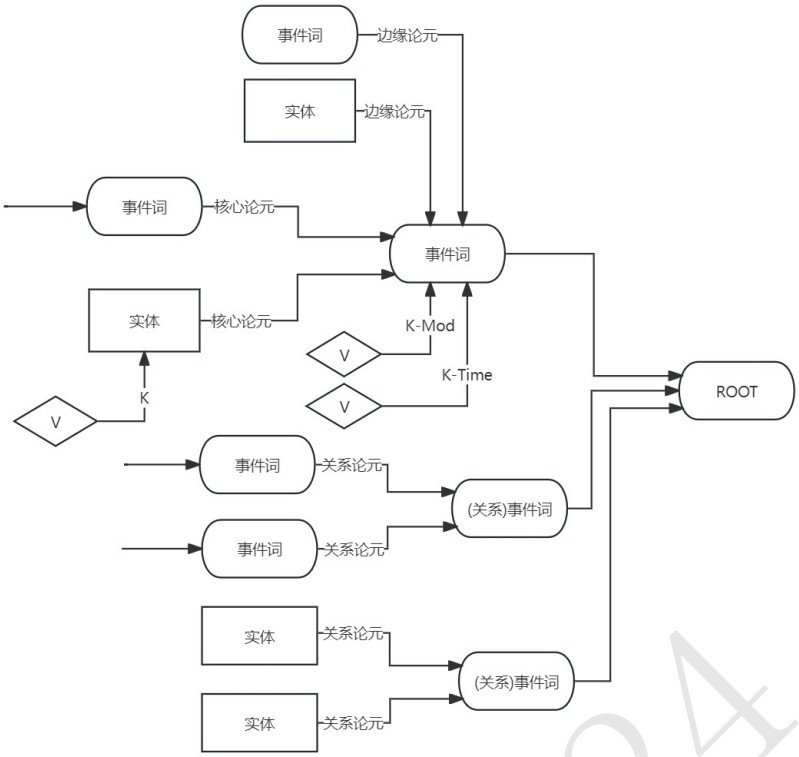



Figure 6: 意合图的抽象表示

### 4.2 三元组集合

图结构可以三元组的形式进行表达，以便进行数据存储与语料标注。一个意合图表征以三元组(W1,Rel,W2)作为基本单元，形式化为三元组集合（如图7）。W1，W2∈N，N为图中节点所承载单元的集合，Rel为边所承载的语义关系的集合。可解读为W1是W2的Rel，或者W1表征了W2的Rel。

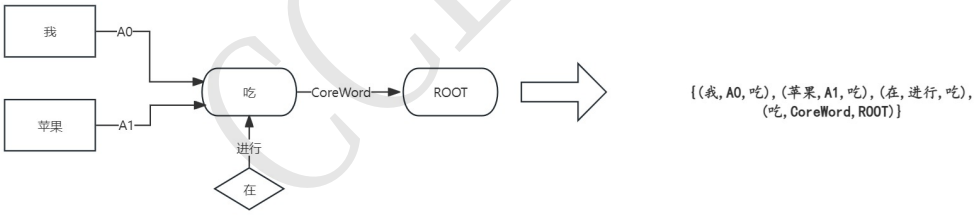



Figure 7: “我在吃苹果”的意合图到三元组转换

## 5 意合图的表示内容

意合图是围绕事件的语义表示，包括事件结构和实体结构两部分。本节先阐述意合图对事件的理解，再从事件结构与实体结构两方面说明意合图的表示内容。

### 5.1 事件的基本概念

事件是现实世界或可能世界中事物的行为、状态，以及事物间存在的关系。人类对事件的理解和认识，通过语言记录下来。意合图从语言的形式表达中挖掘其深层语义信息，将通过语言记录下来的事件形式化地表征出来。意合图所表征的事件可分为一般事件和关系事件两种。

一般事件是对现实世界或可能世界中事物的动作行为或状态描述，可以是动态的事件，如动作行为、发展、变化等；也可以是静态的事件，如性质、状态、判断、关系、心理活动等。

CCL 2024

一般事件内部由论元、事件词、情态信息、时空信息构成。一般事件的事件词通常能够在句中找到对应的语言单元，即显性事件词；在一定语境下事件词可能在语言符号形式上缺省，意合图允许新增节点补全事件词，此时为隐性事件词。

除一般事件外，意合图还将单元的外部关系视作事件，称之为关系事件。我们认为实体间或事件间的关系本身也是客观存在于现实世界或可能世界的一种事实，可认为是一种静态事件，如两个实体间的领属关系、整体部分关系等，两个事件间的因果关系、转折关系等都是或显性或隐性地反映出的事实。关系事件由关系论元和关系事件词构成。关系事件词是抽象出来的概念词语，包括事件间关系词与实体间关系词，为隐性事件词。

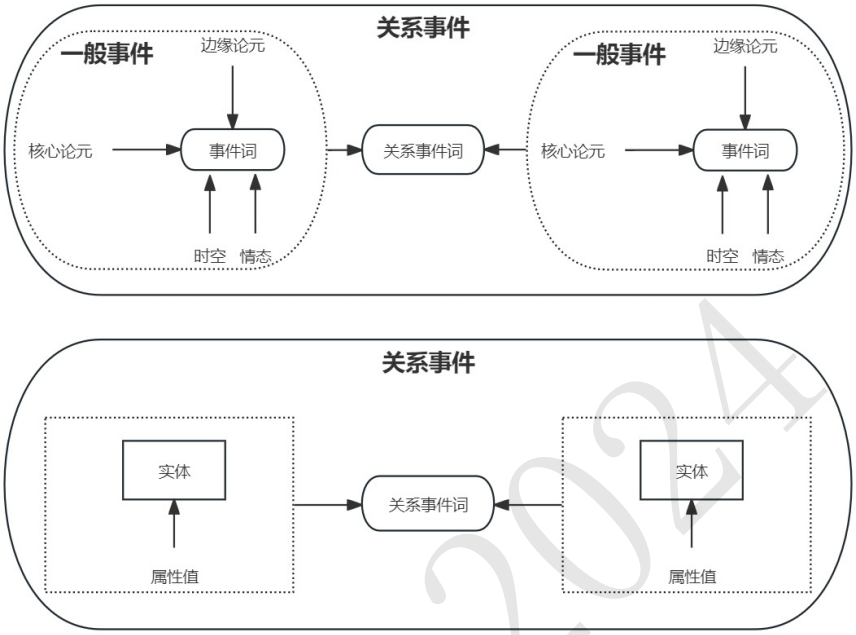



Figure 8: 意合图事件的抽象概念图

## 5.2 事件结构

### 5.2.1 事件内结构

事件内结构包括论元结构、情态结构和时空结构。

（1）论元结构。论元是事件的构成成分，与事件词发生关系，分为核心论元与边缘论元。意合图中将核心论元分为三类，即主体、客体、邻体；边缘论元有九类，分别为：工具、材料、方式、依据、原因、目的、范围、数量、状态。其中，数量又细分为数量、数量源点、数量终点；状态又细分为状态、状态源点、状态终点。此外，由于意合图定义了特殊的关系事件，因此论元中也含有特殊的关系论元，如“因果关系事件”的论元有“原因事件”和“结果事件”两种，我们对每一种关系事件均定义了其对应的关系论元。

（2）情态结构。情态是事件的属性，依附于事件词。从表达情态的词汇载体看，常为副词、助词、形容词等；从句法位置上看，常为状补语；从语义指向上看，一般指向事件词。在意合图中，意合图的情态信息分为八类，分别为：程度、判断、语气、范围、情状、方式、能愿、结果。

（3）时空结构。时空即事件的时间信息和空间信息。事件与实体作为两大认知对象，事件与实体相比，时间性与空间性是其最大的特点之一。事件的时间信息和空间信息是事件最基本的构成成分与属性特征。意合图的时间信息是对事件时间的详细表征。我们认为事件的时间信息并不是由单一的某一时间论元或时态等信息所表达的，如果关注事件的时间信息，就要关注与时间相关的各个方面，进行综合计算，才能真正更确切地反映事件的时间信息。因此，意合图不再将时间信息拆分成离散的多个语义类归入不同的语义结构中，而是作综合处理，划分出时间语义层。意合图的时间信息分为六类，分别为时点、时段、时态、时频、时制、时序。意

第二十三届中国计算语言学大会论文集，第740页-第749页，太原，中国，2024年7月25日至28日。
卷1：主会论文
(c) 2024 中国中文信息学会计算语言学专业委员会

合图的空间信息分为两大类，分别为处所和趋向。其中，处所又分为处所、处所源点、处所终点。

### 5.2.2 事件外结构

事件外结构即事件与事件间的关系，一般事件和一般事件构成关系事件，关系事件和关系事件构成更大的关系事件，一般事件也可以与关系事件构成更大的一般事件或关系事件。构成关系事件的一般事件或关系事件，是关系事件的论元，我们称之为关系论元。

意合图的事件间关系分为八类，分别为时序关系、递进关系、转折关系、因果关系、假设关系、条件关系、目的关系、逻辑关系。其中逻辑关系又分为并列关系和选择关系。

## 5.3 实体结构

实体结构也有内外之分，实体内结构是指实体与依附于实体的属性，实体外结构是指实体与实体间的关系。

### 5.3.1 实体内结构

实体内结构为实体与依附于实体的属性，实体的属性是实体自身结构的一部分。意合图中将实体的属性分为九类，分别为数量、时间、处所、参照、功用、性状、表象、关涉、指别。

### 5.3.2 实体外结构

实体外结构为实体与实体间的关系，构成关系事件[4]。构成关系事件的实体是关系事件的论元，我们称之为关系论元。在一些情况下，事件可以作为整体与实体发生关系，此时该事件与实体间的关系，相当于实体与实体间的关系。意合图的实体间关系分为四类，分别为同指关系、领属关系、整分关系、逻辑关系。其中，逻辑关系又分为并列关系、选择关系。

# 6 结语

本文介绍了一种具有通用性与扩展性的多层次语义表示方法——意合图。首先从语言观与应用观两个维度探讨了意合图的设计理念，其次对意合图多层次语义体系进行了介绍，阐释了意合图在表示精度与构建成本，语义覆盖度与体系复杂度，通用性与扩展性等多个方面取得平衡，同时体现了意合图在未来应用上的潜在价值。

意合图作为一个研究项目，其相关研究仍在不断地推进中。基于意合图构建的语义资源也已投入评测活动，供学界同仁使用。关于意合图语义资源构建的规范与成果将于另文 (郭梦溪等, 2024)详细介绍。

# 参考文献

Bin Li, Yuan Wen, Lijun Bu, Weiguang Qu. 2016. *Annotating the Little Prince with Chinese AMRs.* In Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016), pages 7–15, Berlin, Germany. Association for Computational Linguistics.

Bin Li, Yuan Wen, Li Song, Weiguang Qu, Nianwen Xue. 2019. *Building a Chinese AMR bank with concept and relation alignments.* Linguistic Issues in Language Technology, 18.

Fillmore, Charles J. 1977. *The need for a frame semantics in linguistics.* Statistical Methods in Linguistics, Statistical Methods in Linguistics.

Laura Banarescu, Claire Bonial, et al. 2013. *Abstract Meaning Representation for Sembanking.* In Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Robinson, J. J. 2005. *Dependency structures and transformation rules.* Language, 1970, 46(2):259–285

Collin F. BakerCharles J. FillmoreJohn B. Lowe. 1998. *The Berkeley FrameNet Project.* Proceedings of the 36th Annual Meeting on Association for Computational Linguistics, Canada.

Leonard Talmy. 2012. *Toward a Cognitive Semantics: (Volume I) Concept Structuring Systems.* 外语教学与研究出版社, 北京.

[4]关系事件应属事件结构，但为便于论述，仍将其从单元性质的视角下放于实体结构中说明


(c) 2024 中国中文信息学会计算语言学专业委员会

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. *The Proposition Bank: An Annotated Corpus of Semantic Roles.* Computational Linguistics, 31(1):71–106.

Mingqin Li, Juanzi Li,Zhendong Dong, Zuoying Wang,Dajin Lu. 2003. *Building a large Chinese corpus annotated with semantic dependency.* Proceedings of the Second SIGHAN Workshop on Chinese Language Processing -.

Rosen,S. T. 2003. *The syntactic representation of linguistic events.* The Second Glot International State-of-the-Article Book.

Omri Abend, Ari Rappoport. 2013. *Universal Conceptual Cognitive Annotation (UCCA).* In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 228–238, Sofia, Bulgaria. Association for Computational Linguistics.

Omri Abend, Ari Rappoport. 2013. *UCCA: A Semantics-based Grammatical Annotation Scheme.* In Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers, pages 1–12, Potsdam, Germany. Association for Computational Linguistics.

Omri Abend, Ari Rappoport. 2017. *The State of the Art in Semantic Representation.* In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).

Wanxiang Che, Yunlong Feng, Libo Qin, and Ting Liu. 2021. *N-LTP: An Open-source Neural Language Technology Platform for Chinese.* In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 42–49, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

董振东, 董强, 郝长伶. 2007. 知网的理论发现. 中文信息学报.

郭梦溪, 李梦, 荀恩东, 饶高琦, 于钟洋. 2024. 基于意合图语义理论的结构标注体系与资源建设. 第二十三届中国计算语言学大会.

郝晓燕, 刘伟, 李茹, 刘开瑛. 2007. 汉语框架语义知识库及软件描述体系. 中文信息学报,96-100+138..

黄彤, 李斌, 闫培艺, 戴玉玲, 曲维光. 2020. 基于抽象语义表示的汉语构式标注与分析. 中文信息学报.

蒋炜, 李正华, 张民. 2020. 句法增强的*UCCA*语义分析方法. 北京大学学报(自然科学版).

鲁川. 2001. 汉语语法的意合网络. 商务印书馆, 北京.

刘开瑛, 由丽萍. 2006. 汉语框架语义知识库构建工程. 中文信息处理前沿进展——中国中文信息学会二十五周年学术会议论文集.

刘茂福, 胡慧君. 2017. 基于认知与计算的事件语义学研究. 科学出版社,北京.

刘挺. 2017. 从知识图谱到事理图谱. https://www.leiphone.com/category/ai/Fx6Mgs9WRPBshnIG.html.2017-11-15.

刘宗田, 黄美丽, 周文, 仲兆满, 付剑锋, 单建芳, 智慧来. 2009. 面向事件的本体研究. 计算机科学.

邵田. 2023. 基于大数据的汉语情态义简单状补组块研究. 北京语言大学博士论文.

邵艳秋, 邱立坤, 梁春霞, 毛宁. 2011. 中文语义依存树库构建及自动分析技术. 中国中文信息学会.中国计算语言学研究前沿进展（2009-2011）.

王诚文. 2021. 面向意合图的汉语动词论知识构建研究. 北京语言大学博士论文.

王寅. 2021. 认知语言学. 北京大学出版社, 北京.214

王贵荣. 2023. 意合图事件结构标注及分析研究. 北京语言大学博士论文.

荀恩东. 2023. 自然语言结构计算：意合图理论与技术. 人民邮电出版社.

萧国政. 2020. 语法事件与语义事件——面向人工智能的语言研究. 长江学术.

张牧宇, 秦兵, 刘挺. 2014. 中文篇章级句间语义关系体系及标注. 中文信息学报.

CCL 2024

# 由L2到L1的跨语言激活路径研究——基于词汇识别的ERP数据*

**杨思琴**
中国人民大学文学院
ysq44@outlook.com

**胡美**
清华大学人文学院
hu-m20@mails.tsinghua.edu.cn

**江铭虎**
清华大学人文学院
jiang.mh@tsinghua.edu.cn

## 摘要

本研究运用事件相关电位技术(event-related potentials, ERPs)在不同语言模式下探索二语学习者的二语经验是否会影响母语加工。实验招募了两组中国日语学习者作为被试，分别参与了接近双语模式的短版本实验和接近单语模式的长版本实验。统计结果显示，在短版本实验中，当汉-日同形异义词作为启动词时，语义相关性因素引发的N400波幅差异并不显著,但是引发的LPC波幅差异显著。在长版本实验中，语义相关性因素引发的N400和LPC的波幅都显著。据此，本研究推论，当被试在母语环境下加工母语语义时，二语语义在接近双语模式的短版本实验中被激活并影响母语加工，这种影响仅存在于N400的时间窗口期。但是，在接近单语模式的长版本实验中，二语语义在两个时间窗口里都没有影响母语加工。本研究从语言模式和时间窗口两个维度拓展了对二语经验影响母语加工的认识，对构建高质量的人类语言计算模型和系统具有重要的理论意义和应用价值。

**关键词：** 二语；母语；语言模式；N400；LPC

# The Impact of Second Language Experience on Native Language Processing Across Different Language Modes

**Yang Siqin**
Renmin University of China
School of Liberal Arts
ysq44@outlook.com

**Hu Mei**
Tsinghua University
School of Humanities
hu-m20@mails.tsinghua.edu.cn

**Jiang Minghu**
Tsinghua University
School of Humanities
jiang.mh@tsinghua.edu.cn

## Abstract

This study used the ERPs to investigate whether second language learners' second language experience influences native language processing under different language modes. Two groups of Chinese learners of Japanese were recruited as participants, engaging in both a short version experiment resembling a bilingual mode and a long version experiment resembling a monolingual mode. Statistical results revealed that in the short version experiment, when Chinese-Japanese homographs were used as priming words, the amplitude differences of the N400 component induced by semantic relatedness were not significant, whereas the amplitude differences of the LPC component were significant. In the long version experiment, both N400 and LPC amplitudes induced by semantic relatedness were significant. Based on this, the present study infers that when participants process native language semantics in a native language mode, second language semantics are activated and influence native language processing in a

*本课题获得国家自然科学重点基金项目(62036001)的支持


(c) 2024 中国中文信息学会计算语言学专业委员会

short version of the experiment approaching bilingual mode, with this influence only existing within the time window of N400. However, in the long version of the experiment approaching monolingual mode, second language semantics did not affect native language processing in either of the two time windows.. This study expands the understanding of how second language experience influences native language processing from the perspectives of language modes and processing time windows, which has important theoretical significance and application value for constructing high-quality human language computing models and systems.

**Keywords:** Second language , Native language , Language mode , N400 , LPC

## 1 引言

不同语言在大脑中的交互活动是当下双语脑神经加工机制的研究重点，也是语言认知与计算领域的热点话题。已有研究表明，二语学习者的二语是在母语和二语不断交互过程中所习得(梅磊磊，屈婧，李会玲, 2017)。并且，在二语学习者的二语活动中，母语的影响广泛存在。但是，在母语活动中，二语对母语的影响仍然不明确。本研究引入语言模式作为变量，运用ERP技术从不同时间窗口就二语对母语的影响展开进一步研究，以期为二语学习者的双语脑神经加工机制提供一些新的见解。现有研究发现，母语经验会显著影响二语学习和加工的认知神经机制。具体而言，二语学习者在加工二语时，很可能将母语的语言信息、加工的规则和策略等迁移到第二语言加工上(程凯文，邓颜蕙，颜红梅, 2019; 梅磊磊，屈婧，李会玲, 2017; 张才蕙，叶渐桥，杨静, 2023)。现有的研究通过ERP技术的脑电反应证明了二语任务中母语的影响广泛存在，如母语潜意识的自动翻译机制(Thierry and Wu, 2004; Thierry and Wu, 2007; Wu and Thierry, 2010)和竞争机制(Kerkhofs et al., 2006)等。整体而言，母语对二语加工的影响被各种研究多次验证和反复提及，尤其是对非平衡的二语学习者而言(Horie et al., 2007; Krogh, 2022; Mishra and Singh, 2016; 刘雪丽，倪传斌, 2022; 曲春红, 2019; 肖巍，倪传斌, 2016)

不同于母语对二语加工的影响，二语对母语的影响尚未明确。一方面，一些行为研究(Caramazza and Brones, 1979; Gerard and Scarborough, 1989; Jared and Kroll, 2001)、电生理及神经成像研究(Hoversten et al., 2015; Rodriguez-Fornells et al., 2002)发现，在纯粹的母语环境中，二语对二语学习者在母语环境中的语言产生(Caramazza and Brones, 1979; Gerard and Scarborough, 1989)和理解(Jared and Kroll, 2001)并不会造成影响；另一方面，也有一些证据支持二语也会对在母语环境中的语言加工进行干预(Dijkstra T., 2002; Yang et al., 2021; 杨思琴and 江铭虎, 2016)。

Midgley et al. (2008)研究了在英语和法语语境下跨语言的相似性对英法双语者脑电波幅（N400）的影响。研究结果显示，母语对二语有显著影响，但二语对母语的影响并不如前者显著。在Martin et al. (2009)的研究中，被试在含有威尔士语和英语两种语言的语义启动范式中用他们的一种语言执行字母计数任务。被试脑电波幅（N400）的结果同样显示了不同语言词对之间的语义启动效应，并且这种启动效应不受限于语言类型。此外，在(Yang et al., 2021);杨思琴and 江铭虎 (2016)的ERP研究中，二语在母语语义判断任务中都表现出了对译词自动翻译的现象。整体而言，上述的电生理研究都展现了二语对母语加工的影响。

然而，与此不同的是，Rodriguez-Fornells et al. (2002)以早期平衡的加泰罗尼亚-西班牙双语者为实验对象，用ERP技术研究他们执行词汇决策任务的机制。结果显示，反映被试二语加工的ERP脑电成分没有受到二语词频效应的干扰，意味着他们的二语记忆并没有对母语加工造成影响。在该研究的结果中，缺失的词频干扰效应显示，在语义信息被获取之前，来二语的语言信息就已经被抑制了。Hoversten et al. (2015)在ERP技术的研究那个也得出了与前者相吻合的结论。该研究认为，在西班牙语和英语的双语者中，语言成员的身份信息会比概念信息更早被激活，这让受试者有机会抑制对二语对母语加工的影响。

## 2 研究问题

以上针对二语是否会影响母语加工的分歧，根据文献分析，可能有两方面的原因。其一，

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

各个研究中不同的实验范式所营造的语言模式并不一致，容易导致被试在语言模式的选择上出现差异，这可能是影响实验结果的原因之一。其二，以往二语影响母语的研究结果分析主要关注N400脑电成分，在该问题上只得出了“是/否”的答案。事实上，跨语言互动的加工并不仅局限于N400成分的时间窗口，仅根据该成分的分析结果可能限制了本研究对此的认知。因此，本研究就二语对母语影响的问题加入语言模式与分析的时间窗口进行进一步探究。

语言模式（language mode）是指在特定时段内双语者两种语言的激活状态和加工机制。当一个人处于单语模式时，非目标语言处于静息状态，而目标语言则是完全激活状态，前者不影响后者的加工。反之，当一个人处于双语模式时，那么两种语言都属于激活状态(Grosjean, 2013)，可能会相互干扰。Grosjean (2013)详细阐释了“语言模式”的概念，将语言模式表示为一个以单语模式和双语模式为两端的连续体（如图1）。双语者使用的主导语言被称为“基本语言”（base language），即图中的“语言A”；主导语言之外的语言即图中的“语言B”。随着双语者逐渐向双语模式靠近，语言B的激活程度会不断加深。语言模式已在语言的理解(Dunn, 2014)和生成(Dewaele, 2014)方面本验证。例如，Dunn (2014)的研究表明，对西班牙语和英语的双语者来说，在双语模式下执行词汇决策任务所需要的反应时更长。除此之外，该研究也发现，反应时与非目标语言的激活水平之间存在正相关关系。Dewaele (2014)研究了超过2000名多语者，发现了在双语环境下，两种语言都处于激活状态，多语者更倾向于在两种语言之间转换。鉴于语言模式是双语加工过程中一个不容忽视的变量，本研究推测，二语学习者的语言状态可能会随着在语言模式连续体中所处的位置而转换，进而影响二语对母语加工的作用。

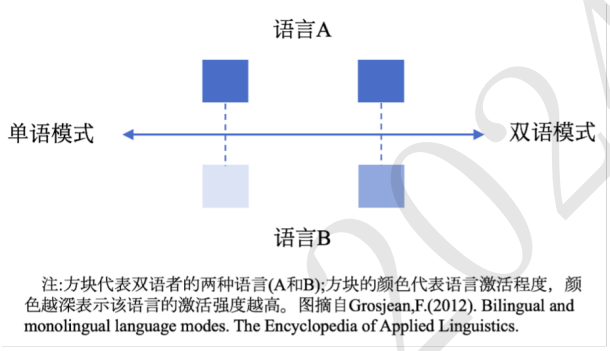



图 1: Grosjean的语言模式示意图

## 3 实验

本研究通过观察二语学习者在加工同形异义词的母语语义时，其二语语义对母语语义加工的干扰情况，来推测二语对母语加工的影响。具体而言，本研究邀请中国日语学习者作为被试，运用ERP技术记录他们在加工汉-日同形异义词和非汉-日同形异义词作为启动词时，目标词所产生的脑电信号差异。通过分析这些脑电信号差异，来推测二语在母语环境下是否会对母语加工造成影响。与此同时，本研究通过对实验呈现的语块进行设计，模拟两种不同语言模式版本的实验。如图2，短版本实验是接近双语模式的版本，包含4个语块，每个语块中都含有汉-日同形异义词；长版本实验是接近单语模式的版本，其在短版本实验的4个语块基础上，于每个语块之间加入了没有汉-日同形异义词的纯汉语语块，加起来一共有8个语块。本研究重点观测中国日语学习者在不同版本的情况下，他们同形异义词的二语语义对母语语义加工的影响。

本研究一共设置了3个变量，包括实验版本（长版本，短版本），词对的语义相关性（相关，无关）和启动词类型（启动词指的是词对的第一个词语）（汉-日同形异义词，非汉-日同形异义词）。其中，词对的语义相关性和启动词类型这两因素的两个水平2×2组成4个条件（Semantic：S+代表语义相关，S-代表语义无关；Homograph：H+代表启动词是汉-日同形异义词，H-代表启动词是纯汉语词语；Version：V+代表短版本实验，V-代表长版本实验），如表1所示。被试的实验任务是判断前后呈现的词对的汉语语义相关性是否相关，即语义判断任务。

N400是在刺激呈现后400 ms左右于大脑皮层中央和顶部区域出现的一种负波。它可以反映大脑在该阶段对词汇语义的加工。例如，语义无关的词对通常会比语义相关的词对诱发一个更

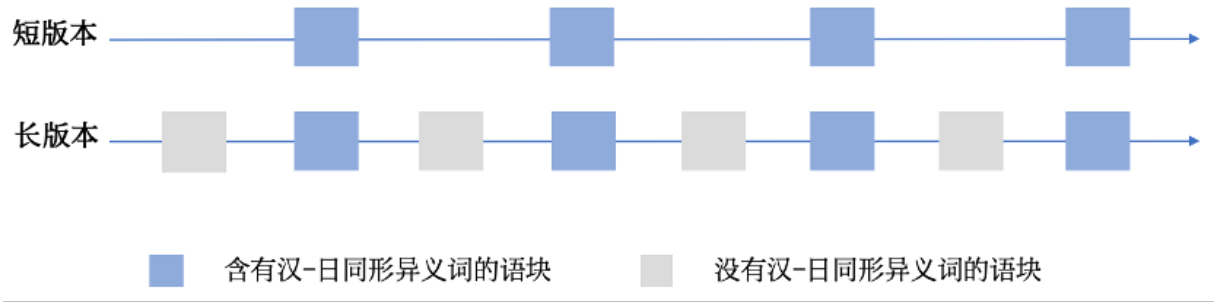



图 2: 短版本和长版本实验的语块示例

表 1: 实验材料范例

| | 汉语语义相关 (S+) | | | | 汉语语义相关 (S-) | | | |
|---|---|---|---|---|---|---|---|---|
| | 日语语义无关 | | | | 日语语义相关 | | | |
| | 启动词 | 目标词 | 启动词 | 目标词 | 启动词 | 目标词 | 启动词 | 目标词 |
| 汉-日同形异义词(H+) | 人参<br>人参日义为胡萝卜<br>S+H+ | 草药 | | | 人参<br>人参<br>S-H+ | 橙色 | | |
| 非汉-日同形异义词(H-) | | | 重要<br>S+H- | 关键 | | | 重要<br>S-H- | 蛋糕 |

大的N400波幅(Thierry and Wu, 2007)。在本研究中，如果以中日同形异义词为启动词的语义无关词对（S-H+）诱发的N400效应与语义相关词对（S+H+）诱发的N400效应差异不显著，则说明被试在该阶段激活了二语语义并对母语加工产生了影响，反之则不然。此外，于类似脑区出现在600ms左右的晚期正波(Late Positive Component, LPC) 也可以作为反映跨语言影响在N400后期活跃程度的指标。在跨语言影响的ERP研究中，Martin et al. (2009)和Hoshino and Thierry (2012)曾将反映语言后期加工的LPC成分纳入分析范围。借鉴于此，本研究也将通过LPC效应来探讨二语学习者在加工母语时后期二语的受影响情况。以往研究表明，语义相关的词对通常会比语义无关的词对诱发一个更大的LPC波幅。在本研究中，如果以中日同形异义词为启动词的语义无关词对（S-H+）诱发的LPC效应与语义相关词对（S+H+）诱发的LPC效应差异不显著，则说明被试在该阶段依然激活了二语语义并对母语加工产生影响，反之则不然。

### 3.1 实验语料

短版本实验使用了180个词语作为实验材料。其中，30个中日同形异义词和30个汉语词语分别充当了两组词对的启动词。换言之，这60个词语在呈现的词对中出现两次，分别与余下的120个词语配对，一共组成了120个汉语词对。其中，同一个启动词与汉语相关的目标词配对一次，与汉语不相关的目标词配对一次。在每个语块中，实验材料的类型和数量一致。在长版本实验中，本研究额外增加了4个没有同形异义词的语块，在这些语块中，实验材料的类型和数量也一致。由于额外增加的语块仅充当调节语言模式的作用，其结果不进入研究统计分析环节，因此，这部分实验材料的词频等因素也未纳入统计范围之内。

实验材料的词频数据均来自公开的词频库(Cai and Brysbaert, 2010)。由于汉语词汇以成对的形式出现，所以，在控制词汇词频时，本研究也将词汇呈现的顺序（简称为词序）作为一个因素纳入分析。统计结果显示，语义相关性、启动词类型和词序它们之间的交互效应和各自的主效应都不显著[F>0.03, ps>0.8]。总的来看，实验严格控制了实验材料中词汇的词频。为了对汉语词对的语义相关性进行控制，实验招募了15名完全不懂日语的大学生使用5点李克特量表对汉语词对的语义相关程度进行了评估。其中，5 分代表语义相关度最高，1分代表语义相关度


(c) 2024 中国中文信息学会计算语言学专业委员会

最低。收集完调研数据后，笔者将语义相关性和启动词类型作为两因素进行统计分析。结果显示，语义相关性和启动词类型不存在交互效应[F（1, 29）= 2.098, p = 0.158]，只有语义相关性差异显著[F（1, 29）= 11292.582, p ¡ 0.001, $\eta_p^2$ = 0.997]，这说明无论是在有汉-日同形异义词的词对中，还是在纯汉语词对中，语义相关性的变量得到了合理的控制。启动词的效应不显著[F（1, 29）= 0.081, p = 0.778]，这说明完全不懂日语的被试无法觉察出汉-日同形异义词与纯汉语的差异。在刺激材料呈现的过程中，所有的词对均以伪随机的顺序呈现。

### 3.2 被试

本研究招募了两组中国日语学习者参与实验。每组有效数据的被试均为16人。他们通过了代表日语中高水平的N2考试。他们的年龄范围均在18岁～27岁之间。其中，女性的平均年龄是22.2岁，男性的平均年龄是21.25岁。两组被试在语言水平和年龄两方面没有显著差异（ps ¿0.2）。根据爱丁堡的惯性测试，参加实验的被试均为右利手(Oldfield, 1971)。另外，在实验之前，所有被试的视力或矫正视力都正常。被试在参加实验时身体健康，并且没有任何神经或精神疾病的病史。所有被试均获得了《赫尔辛基宣言》的知情同意。

### 3.3 实验流程

被试坐在离电脑屏幕80 cm左右的椅子上完成操作。实验运用心理学软件工具E-Prime 2.0来设计任务和呈现词语，任务要求也呈现在屏幕上。屏幕的背景为灰色。作为实验材料的词语均以白色呈现，字体都是34号的宋体。被试被告知实验的具体操作任务：观察先后呈现的一对汉语词语，让他们判断词对中的词是否具有相关性，并及时按下游戏手柄的按钮进行反应。整个实验分为两个模块。首先是练习模块，屏幕向被试呈现了20对汉语词对，其中一半语义相关，一半语义不相关。被试在练习模块中，当正确率达到90%及以上后才开始进入接下来的正式实验。数据分析仅基于被试在实验模块的反应。每个试次的流程如下所示：首先，一个提示符号“+”呈现在屏幕中央，持续时间为200ms，提醒被试集中注意力；其次，作为启动词的汉语词语出现在屏幕中央，停留时间为500ms；之后是一个空白屏幕，停留的时长是500ms、600ms或700ms中的一个；接着，作为目标词的汉语词语出现在屏幕上，直到被试对此作出判断后才消失；然后，空白屏幕再次出现，此时，空白屏幕停留的时间是随机的，可能停留200ms，也可能停留300ms或400ms。到此为止，一个实验任务便完成了。紧接着下一个同样流程的实验任务便开始了。

### 3.4 数据采集与分析

实验使用62 Ag/AgCl 电极弹性帽来记录被试在参加实验过程中的脑电数据。该电极帽上配备着国际10-20电极放置系统（Easycap；Brain Products GmbH，Gilching，Germany）。设备中的FCz和AFz电极分别用作参考电极和接地电极。两个眼电极用于测量垂直和水平眼电图，分别放置在左眼下方和右眼外侧（Picton et al., 2000）。在准备正式实验前，电极阻抗保持在10kΩ以下。在Brain Vision Recorder软件（Brain Products，Munich，Germany）的控制下，研究使用带通为0.01-100Hz的BrainAmpDC放大器系统（Brain Products GmbH）记录被试的EEG数据。电生理记录的过程严格遵循以往已发表并被公认的流程(Yang et al., 2021))。

在收集完数据之后，本研究使用MATLAB中的EEGLAB插件对其进行了预处理。在本实验中，本研究以双侧乳突为参考电极点（TP9、TP10），在此基础上对脑电的波幅进行计算。同时，本研究用EEGLAB中的ERPLAB插件在0.1-30Hz的范围内进行了滤波，并通过独立成分分析（ICA）去除了眨眼、眼漂、肌肉动作等因素产生的伪迹。在记录完毕后，本研究将连续的EEG信号切割为若干个时长为1100ms的分段（从关键词出现前100ms到关键词出现后的1000ms），并在分段开始的时刻进行了基线校正。最后，本研究将振幅大于+/- 100 μV的分段剔除，被剔除的分段占比超过20%的被试数据被排除在分析之外。本研究通过使用从刺激开始前100毫秒至目标出现（0ms）的范围内的平均信号，进行了基线校正。根据N400和LPC在脑电波形图上的时间分布和在脑电地形图中的空间分布，本研究选取了Cz、C1、C2、CPz、CP1和CP2这6个电极点的平均波幅作为两个脑电成分的分析对象。其中，N400电位被定义为目标词呈现后300-450ms的平均振幅，LPC被定义为目标词呈现后500-750ms的平均振幅。

本研究使用重复测量方差分析对脑电波的结果进行统计。其中，以语义相关性（相关、不相关）和启动词类型（汉-日同形异义词，非汉-日同形异义词）作为组内效应，以实验版本


(c) 2024 中国中文信息学会计算语言学专业委员会

（长版本、短版本）作为组间效应。当自由度大于1时，则采用Greenhouse-Geisser 法校正。如果交互作用显著，则继续进行简单效应分析或t检验。

### 3.5 实验结果

如表2所示，被试行为实验结果中的正确率均在87%以上，保证了数据的有效性。短版本和长版本在6个电极点上的N400和LPC，以及它们各自的脑电地形图差异分别如图3和图4所示。

表 2: 行为实验结果的正确率

| 组别 | 条件 | | 平均值 | 标准误差 | 95% 置信区间 | |
|---|---|---|---|---|---|---|
| | | | | | 下限 | 上限 |
| 短版本 | S+ | H+ | .938 | .016 | .905 | .970 |
| | | H- | .969 | .010 | .948 | .990 |
| | S- | H+ | .879 | .020 | .839 | .920 |
| | | H- | .969 | .007 | .955 | .982 |
| 长版本 | S+ | H+ | .922 | .016 | .890 | .954 |
| | | H- | .953 | .010 | .932 | .975 |
| | S- | H+ | .982 | .020 | .941 | 1.022 |
| | | H- | .980 | .007 | .966 | .994 |

N400的统计结果显示，启动词类型、语义相关性和版本的三因素交互效应略微显著[F（1, 30）= 3.605, p = 0.067，$\eta_p^2$ = 0.107]。在短版本中，启动词类型和语义相关性的两因素交互效应显著[F（1, 15）= 11.468, p = 0.004，$\eta$ 2p = 0.433]。其中，它们简单效应分析结果显示，在以汉-日同形异义词作为启动词的情况下，语义相关性条件引发的差异并不显著[F（1, 15）= 1.94, p = 0.184]。但是，在以非汉-日同形异义词作为启动词的情况下，语义无关的情况引发的波幅（M = 0.165, SE = 0.646）比语义相关的情况引发的波幅（M = 3.501, SE = 0.797）更负[F（1, 15）= 26.77, p =<0.001]。

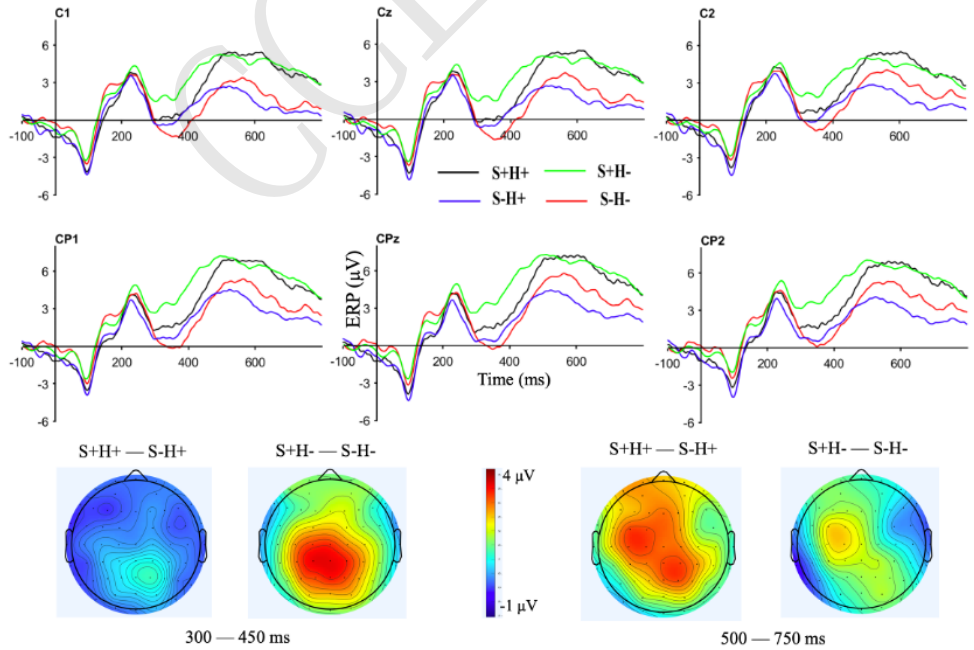



图 3: 短版本在6个电极点上的N400、LPC及它们各自的脑电地形图差异


(c) 2024 中国中文信息学会计算语言学专业委员会

在长版本中，启动词类型和语义相关性的两因素交互效应略微显著[F（1, 15）= 4.081, p = 0.062，$\eta_p^2$ = 0.214]。其中，它们简单效应分析结果显示，在以汉-日同形异义词作为启动词的情况下，语义无关的情况引发的波幅（M = -1.545, SE = 0.755）比语义相关的情况引发的波幅（M = -0.387, SE =0.720）更负[F（1, 15）= 9.05, p = 0.009]。类似的，在以非汉-日同形异义词作为启动词的情况下，语义无关的情况引发的波幅（M = -1.487, SE = -0.546）比语义相关的情况引发的波幅（M = -0.597, SE = 0.797）更负[F（1, 15）= 22.66, p =<0.001]。

总体来看，在N400时间窗口内，短版本实验中的汉-日同形异义词的日语语义可能已被激活，激活的二语语义影响了被试对母语的加工。然而，在长版本实验中，汉-日同形异义词的日语语义没有被激活，母语的加工也没有受到其影响。

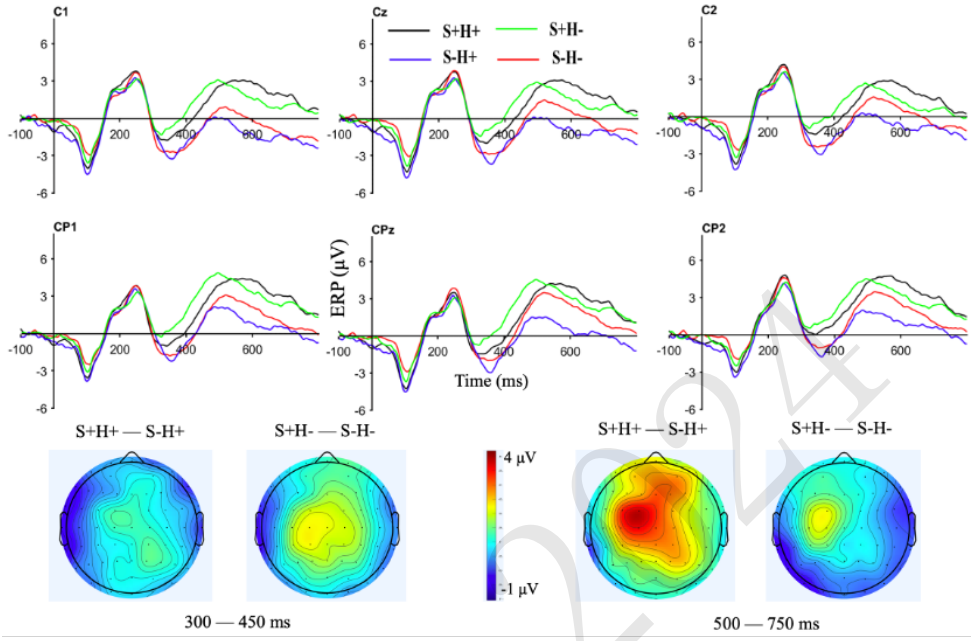



图 4: 长版本在6个电极点上的N400、LPC及它们各自的脑电地形图差异

LPC的统计结果显示，启动词类型和语义相关性的交互效应显著[F（1, 30）= 5.020, p = 0.033，$\eta_p^2$ = 0.143]。其中，它们的简单效应分析结果显示，在以汉-日同形异义词作为启动词的情况下，语义无关的情况引发的波幅（M = 1.027, SE = 0.458）比语义相关的情况引发的波幅（M = 3.819, SE = 0.456）更负[F（1, 15）= 16.18, p = 0.001]。类似的，在以非汉-日同形异义词作为启动词的情况下，语义无关的情况引发的波幅（M = 1.901, SE = 0.404）比语义相关的情况引发的波幅（M = 3.438, SE = 0.473）更负[F（1, 15）= 7.86, p =0.013]。但是，无论是在语义相关的情况下[F（1, 15）= 0.01, p = 0.932]还是语义无关的情况下[F（1, 15）= 2.09, p = 0.169]，启动词类型中的两水平均不显著。此外，组间因素版本的主效应显著[F（1, 30）= 14.716, p = 0.001，$\eta_p^2$ = 0.329]，其中，短版本的波幅（M = 3.709 , SE = 0.429）比长版本的波幅（M = 1.384, SE = 0.429）更正。

总体来看，在LPC时间窗口内，短版本实验中的汉-日同形异义词的日语语义可能已被抑制，因此没有影响到母语的后期加工。然而，在长版本实验中，汉-日同形异义词的日语语义可能跟在N400窗口一样，并没有被激活。

## 4 讨论

本研究运用ERP技术在不同语言模式下就二语是否会影响母语加工这一话题进行探究。实验招募了中国日语学习者作为被试，统计并分析了被试分别在N400和LPC两个时间窗口的脑电数据。N400的统计结果显示，在短版本实验中，当汉-日同形异义词作为启动词时，语义相关性条件引发的差异并不显著。这暗示着汉-日同形异义词的日语语义可能已被激活并影响了母语的加工。然而，在长版本实验中，同样情况下的语义相关性引发的差异却显著。这表明在该情况下，汉-日同形异义词的日语语义可能没有被激活，母语加工也并未受其影


(c) 2024 中国中文信息学会计算语言学专业委员会

响。然而，在LPC的统计结果中，实验版本因素并没有与语义相关性或启动词因素有任何交互效应。并且，启动词类型的两水平之间也没有显著差异。综上，本研究推论，当被试在母语模式下加工母语语义时，二语语义在接近双语模式的短版本实验中会被激活并影响母语加工，但这种激活和影响仅存在于N400的时间窗口期，并不会持续到LPC的时间窗口期。但是，在接近单语模式的长版本实验中，二语语义在两个时间窗口里都没有被激活的脑电现象。本研究根据实验结果分别从语言模式、双语交互激活模型+（Bilingual Interactive Activation Model，BIA+）(Dijkstra T., 2002)以及二语学习对母语加工认知神经机制的影响三个方面展开分析讨论。首先，脑电实验结果在接近双语模式的短版本实验和接近单语模式的长版本实验之间的差异，印证了Grosjean (2013)提出的“语言模式”中的激活效应对跨语言影响的作用。这种差异主要体现在汉-日同形异义词的二语的语义启动上。具体而言，在短版本实验中，作为启动词的汉-日同形异义词的二语语义被激活，影响了原本在母语语义背景下是语义无关词对（S-H+）的加工。这契合了“语言模式”对双语模式下双语都激活的描述。相反，在长版本实验中，作为启动词的汉-日同形异义词的二语语义没有被激活，也没有影响母语语义的加工。这契合了单语模式下非目标语言处于静息状态的陈述。本研究设计的短版本实验和长版本实验，合理地模拟了Grosjean (2013)所描述的单语和双语模式的情况

其次，本研究的结果验证了BIA+对跨语言影响中激活与抑制的描述，同时也补充了其在语言模式上的缺憾。基于同形异义词识别提出的双语词汇通达模型——BIA+认为，词汇识别不仅受到另一种语言词汇书写形式的影响，还受到跨语言的语义表征的影响。在本研究的结果中，汉-日同形异义词在母语语境下的语义相关性判断也受到了日语语境下的语义的影响。这正契合了BIA+对同形异义词跨语言影响的描述。此外，BIA+有明确地交代它所适用的语言模式，而本研究恰好在语言模式上补充了这一内容，即BIA+描述的跨语言影响更适用于接近双语模式的语境。

最后，本研究的脑电结果与以往描述二语学习影响母语加工认知神经机制的功能性核磁共振研究（functional magnetic resonance imaging，fMRI）结果相吻合。梅磊磊，屈婧，李会玲(2017)发现母语加工与二语学习之间的相互作用并不是单向的，第二语言学习也会反过来影响母语加工的认知神经机制。Mei et al. (2014)运用fMRI技术发现，在二语学习的过程中，二语以语义为中介，对母语加工产生了影响。本研究也是在执行语义判断任务中，发现了从二语到母语方向的跨语言影响。母语与二语之间是否存在更直接的相互影响，未来可以采用非语义相关性识别或判断任务来进一步探索该问题。

本研究的成果可以为建立更精确、更人性化的语言表征和计算模型提供依据。例如，如果发现双语者在某些情况下更倾向于混合使用两种语言的语义或语法特征，那么计算机语言模型就需要考虑这种影响。与此同时，对语言模式这一变量的关注可以帮助优化自然语言处理系统在处理多语言输入时的性能。比如，系统可以根据用户的语言背景动态调整语言模型，以提高准确性。这些对构建高质量的人类语言计算模型具有重要的理论意义和应用价值。

## 5 结论

本研究通过ERP技术在不同语言模式下研究了二语学习者对母语加工的影响。结果表明，在接近双语模式的短版本实验中，二语语义在N400时间窗口期间被激活，影响了该阶段对母语的加工；而它在接近单语模式的长版本实验中没有被激活，母语加工也没有受其影响。这表明二语经验对母语加工的影响情况受到语言模式和加工时间窗口的调节。未来的研究可以扩展到考虑更多种类的语言模式和其他语言间的影响，同时关注个体差异并考虑长期学习效应，以深入了解二语经验对母语加工的全面影响。

## 参考文献

Q. Cai and M Brysbaert. 2010. Subtlex-ch: Chinese word and character frequencies based on film subtitles. *Plos One*, 5(6):e10729.

A. Caramazza and I. . Brones. 1979. Lexical access in bilinguals. *Bulletin of the Psychonomic Society*, 13(4):212–214.

Jean-Marc Dewaele. 2014. Intra- and inter-individual variation in self-reported code-switching patterns of adult multilinguals. *International Journal of Multilingualism*, 11.


(c) 2024 中国中文信息学会计算语言学专业委员会

van Heuven W J B. Dijkstra T. 2002. The architecture of the bilingual word recognition system: From identification to decision. *Bilingualism: Language and Cognition*, 5(3):175–197.

J. E Dunn, A. L. Fox Tree. 2014. More on language mode. *International Journal of Bilingualism*, 18(6):605–613.

L. D Gerard and D. L Scarborough. 1989. Language-specific lexical access of homographs by bilinguals. *Journal of Experimental Psychology Learning*, 15(2):305–315.

François Grosjean, 2013. *Bilingual and Monolingual Language Modes*.

Kaoru Horie, Hyeonjeong Jeong, Ryuta Kawashima, Yuko Sassa, Shigeru Sato, Motoaki Sugiura, Masato Taira, and Satoru Yokoyama. 2007. Cross-linguistic influence on brain activation during second language processing: An fmri study. *Bilingualism: Language and Cognition*, 10(2):175–187.

N. Hoshino and G. Thierry. 2012. Do spanish-english bilinguals have their fingers in two pies - or is it their toes? an electrophysiological investigation of semantic access in bilinguals. *Front Psychol*, 3:9.

L. J. Hoversten, T. Brothers, T. Y. Swaab, and M. J. Traxler. 2015. Language membership identification precedes semantic access: Suppression during bilingual word recognition. *J Cogn Neurosci*, 27(11):2108–16.

Debra Jared and Judith F. Kroll. 2001. Do bilinguals activate phonological representations in one or both of their languages when naming words? *Journal of Memory and Language*, 44(1):2–31.

R. Kerkhofs, T. Dijkstra, D. J. Chwilla, and E. R. de Bruijn. 2006. Testing a model for bilingual semantic priming with interlingual homographs: Rt and n400 effects. *Brain Res*, 1068(1):170–83.

Simone Møller Krogh. 2022. Danish-english bilinguals' cognate processing in l1 and l2 visual lexical decision tasks. *Languages*, 7(3).

C. D. Martin, B. Dering, E. M. Thomas, and G. Thierry. 2009. Brain potentials reveal semantic priming in both the 'active' and the 'non-attended' language of early bilinguals. *Neuroimage*, 47(1):326–33.

Leilei Mei, Gui Xue, Zhong-Lin Lu, Chuansheng Chen, Mingxia Zhang, Qinghua He, Miao Wei, and Qi Dong. 2014. Learning to read words in a new language shapes the neural organization of the prior languages. *Neuropsychologia*, 65:156–168.

K. J. Midgley, P. J. Holcomb, W. J. VanHeuven, and J. Grainger. 2008. An electrophysiological investigation of cross-language effects of orthographic neighborhood. *Brain Res*, 1246:123–35.

Ramesh Kumar Mishra and Niharika Singh. 2016. The influence of second language proficiency on bilingual parallel language activation in hindi-english bilinguals. *Journal of Cognitive Psychology*, 28(4):396–411.

Richard C Oldfield. 1971. The assessment and analysis of handedness: The edinburgh inventory. *Neuropsychologia*, 9(1):97–113.

A. Rodriguez-Fornells, M. Rotte, H J. Heinze, and T NoÈsselt. 2002. Brain potential and functional mri evidence for how to handle two languages with one brain. *Nature*, 415:1026–1029.

G. Thierry and Y. J. Wu. 2004. Electrophysiological evidence for language interference in late bilinguals. *Neuroreport*, 15(10):1555–8.

G. Thierry and Y. J. Wu. 2007. Brain potentials reveal unconscious translation during foreign-language comprehension. *Proceedings of the National Academy of Sciences of the United States of America*, 104(30):12530–12535.

Y. J. Wu and G. Thierry. 2010. Chinese-english bilinguals reading english hear chinese. *Journal of Neuroscience*, 30(22):7646–51.

S Q. Yang, X C. Zhang, and M H. Jiang. 2021. Bilingual brains learn to use l2 alliterations covertly like poets: Brain erp evidence. *Frontiers in Psychology*, 12:e691846.

刘雪丽，倪传斌. 2022. 二语对一语的影响——来自一语语音加工的眼动证据. 外语与外语教学, (1):34–45.

张才蕙，叶漸桥，杨静. 2023. 汉语作为第二语言学习的脑机制. 心理科学进展, 31(5):747–758.


(c) 2024 中国中文信息学会计算语言学专业委员会

曲春红. 2019. 二语词汇水平对汉英双语者词汇通达中一语自动激活的影响研究. 外语教学与研究, 51(4):560–571.

杨思琴and 江铭虎. 2016. 双语者加工汉语母语语义时对英语的erp激活效应的研究. 中文信息学报, 30(6):9.

梅磊磊，屈婧，李会玲. 2017. 第二语言学习的认知神经机制. 华南师范大学学报·社会科学版, (6):63–73.

程凯文，邓颜蕙，颜红梅. 2019. 第二语言学习与脑可塑性. 心理科学进展, 27(2):209–220.

肖巍，倪传斌. 2016. 二语词汇加工中的一语自动激活:来自中国英语学习者的证据. 外语教学与研究, 48(2):236–248.

CCL 2024

# 大语言模型故事理解能力评价数据集

闫国航[1,†]，郭亚鑫[1,†]，谭红叶[1,2,⋆]，张虎[1]
[1]山西大学 计算机与信息技术学院，山西 太原 030006
[2]山西大学 计算智能与中文信息处理教育部重点实验室，山西 太原 030006
{202112407002, 202222407055}@email.sxu.edu.cn
{tanhongye,zhanghu}@sxu.edu.cn

## 摘要

故事包含大量的社会、物理等常识，同时蕴含深刻的道理，是知识传播、文化传承、价值塑造的重要载体。故事理解是NLP中的一项重要任务。近几年，研究者对大语言模型（LLMs）的语言理解能力进行了很多评估与分析，但由于现有的故事理解数据集大多为答案出现在原文的实体类问题，因此对LLMs故事理解能力的评价与分析非常有限。为此，本文构建了一个寓言故事理解数据集CRMUS，并基于人类故事理解的认知过程：先进行常识推理，然后理解故事寓意，设计了两个任务来评价模型的相应能力。基于CSMUS数据集，我们对多个代表性的LLMs进行了评估，发现：LLMs已经可以较好地理解故事中的常识并进行推理，但在理解故事寓意方面还存在很大提升空间。此外，我们使用项目反应理论（IRT）对数据集进行了质量分析，表明该数据集是高质量的，可以有效地评估LLMs。

**关键词：** 故事理解 ； 常识推理 ； 数据集 ； LLMs ； IRT

# Benchmarking story comprehension ability of large language model

**Guohang Yan[1,†], Yaxin Guo[1,†], Hongye Tan[1,2,⋆], Hu Zhang[1]**
[1]School of Computer and Information Technology, Shanxi University, Taiyuan, Shanxi 030006, China
[2]Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan, Shanxi 030006,China
{202112407002, 202222407055}@email.sxu.edu.cn
{tanhongye,zhanghu}@sxu.edu.cn

## Abstract

Stories contain a lot of social, physical and other commonsense, which is an important carrier of knowledge transmission, cultural inheritance and value shaping. Story comprehension is an important task in NLP. In recent years, researchers have witnessed extensive evaluation and analysis of Large Language Models' (LLMs) comprehension abilities. These assessments have been largely confined due to the predominant focus on entity-based questions within existing story comprehension datasets. Consequently, the scope for evaluating LLMs' comprehension of narratives has been severely restricted. Therefore, we propose CRMUS, a new fable story comprehension dataset. Based on the cognitive process of human story comprehension, we design two tasks to evaluate the model's ability in commonsense reasoning and allegorical understanding. Utilizing the

† Equal contribution
⋆ Corresponding Author

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

**项目资助：** 国家重点研发计划项目(2020AAA0106100) 、国家自然科学基金面上项目(62076155)

CSMUS dataset, we evaluated several representative LLMs and found that the LLMs can understand the commonsense of the story and reason effectively, but there is still much room for improvement in understanding the morals of the story. Furthermore, we performed a quality analysis of the dataset using item response theory (IRT), showing that the dataset is of high quality and can effectively evaluate LLMs.

**Keywords:** Story Comprehension , Commonsense Reasoning , Dataset , LLMs , IRT

## 1 引言

故事是教育领域中的一种重要阅读材料，通常包含丰富的知识、生动的情节与鲜明的人物形象，同时还蕴含着深刻的道理，是知识传播、文化传承、价值塑造的重要载体。故事理解是自然语言处理（Natural Language Processing，NLP）领域的一项经典任务，在上世纪七十年代就受到了研究者的关注(Charniak, 1972; Winograd, 1972; Stede, 1996; Schubert and Hwang, 2000)。近几年，大语言模型（Large Language Model，LLMs）在NLP领域取得了很大进展，研究者利用SQuAD2.0(Rajpurkar et al., 2018)、RACE(Xu et al., 2017)等机器阅读理解数据集对LLMs的语言理解能力进行了评估与分析(Kocon et al., 2023)，发现其在相应数据集上已经具备类人的语言理解与文本推理能力，然而目前尚未对LLMs的故事理解能力进行系统评估。

相较于一般的机器阅读理解，故事理解不仅需要模型根据社会、物理等常识理解局部情节，还需要分析故事人物、意图和行为之间的关系，推断故事背后传递的深刻道理(Tomasulo et al., 2012; Pelletier and Beatty, 2015; Dorfman and Brewer, 1994)。因此，故事理解也是评估LLMs认知能力的一项重要任务。然而，目前的故事理解数据集大多为答案出现在原文的实体类问题，缺乏足够的挑战性。此外，这些数据集主要关注故事局部的、表层的语义理解，不能有效评估模型的故事深层理解能力。表 1展示了NarrativeQA(Kociský et al., 2018)和FairytaleQA(Xu et al., 2022)数据集中的示例。这两个示例均为故事人物方面的问题，答案“詹姆斯·麦迪逊”、“玛奇”都分别在故事原文中直接出现，且通过简单的词形匹配就能回答。这些问题对LLMs的故事理解能力进行的评价与分析非常有限。

| 数据集 | 样例（英文） | 样例（中文） |
|---|---|---|
| NarrativeQA | **Story snippet:** The United States of America, you will recall, was the style by which the rebellious colonies referred to themselves, in the Declaration of Philadelphia. The James Madison who is mentioned as the current President of the United States is now living, in exile, in Switzerland. His alleged predecessor in office...<br>**Question:** Who is in exile in Switzerland?<br>**Answer:** James Madison | **故事片段**：你会记得，美利坚合众国是反叛的殖民地在《费城宣言》中所称的风格。被提及为美国现任总统的詹姆斯·麦迪逊目前流亡瑞士。他所谓的前任.....<br>**问题**：故事中谁在瑞士流亡?<br>**答案**：詹姆斯·麦迪逊 |
| FairytaleQA | **Story snippet:** it so happened that the great man was walking in his garden with his daughter madge that morning, and was feeling in an especially happy mood, so that when he suddenly looked up and saw a little boy before him, he said, kindly...<br>**Question:** what was the name of the man's daughter?<br>**Answer:** madge | **故事片段**：碰巧那天早上，这个人正和他的女儿玛奇在他的花园里散步，心情特别愉快，所以当他突然抬头看到面前有一个小男孩时，他和蔼地说.....<br>**问题**：那个男人的女儿叫什么名字?<br>**答案**：玛奇 |

Table 1: NarrativeQA和FairytaleQA数据集中的示例


(c) 2024 中国中文信息学会计算语言学专业委员会

为此，本文构建了一个新的具有挑战性的故事理解数据集CRMUS，并基于人类理解故事的认知过程，设计了两个任务：常识问答、寓意理解。其中常识问答涉及到的常识类型有：社会常识、生物常识、时间常识、空间常识和物理常识。上述两个任务均采用多项选择问答的形式，每个问题包含四个选项。为了增加混淆性，我们要求标注者提供的错误选项要与正确答案存在一定的语义关联。此外，在寓意理解任务中，除了由标注者提供的人工风格选项之外，还利用LLMs生成机器风格的选项，进一步增加选项的混淆性。具体来说，该数据集包含436篇经典寓言故事和3400个选择题，用于探测模型常识推理与寓意理解能力。

基于CSMUS数据集，本文设置了提示学习和微调两个实验，对多个LLMs进行了评估。具体来说，提示学习实验中，通过调用多个商用LLMs的API进行零样本提示；在微调实验中，通过微调开源的LLMs并进行提示。实验结果表明，现有的LLMs可以较好地理解故事情节并根据常识知识进行推理（最佳结果83.45%），但在深入理解故事寓意方面仍然存在一定困难（最佳结果67.33%）。此外，我们还使用项目反应理论（Item Response Theory，IRT）(Baker, 1985)对数据集进行了质量分析，并对上述LLMs的测试结果进行了排名。结果表明，我们的数据集是高质量的，可以有效地评估LLMs。

我们的数据集和代码可以在http://cuge.baai.ac.cn/#/ccl/2024/crmus 网站公开获取。

## 2 相关工作

### 2.1 故事理解数据集

随着深度学习与NLP技术的进步，机器阅读理解成为一项研究热点。研究者推出了很多的阅读理解数据集，其中包含一些与故事理解相关的数据集，具体来说有以下几类。

**多项选择问答数据集** 该类数据集问题形式均为多项选择题形式。例如，微软公司推出的MCTest(Richardson et al., 2013)数据集，该数据集基于虚构的故事创建，但仅涉及7岁儿童可理解的概念和词。Facebook公司基于儿童故事书构建了CBT数据集(Hill et al., 2016)，但由于采用自动构建方式，数据集质量难以保证，存在很大噪声。上述数据集中的问题都比较简单，仅根据文本表面信息就能回答(Kaushik and Lipton, 2018)。

**生成式问答数据集** 该类数据集要求模型生成问题答案，例如，基于长篇书籍和电影剧本众包构建的NarrativeQA数据集(Kociský et al., 2018)、基于幼儿和青少年童话故事由专家构建的FairytaleQA数据集(Xu et al., 2022)。这些数据集大多为实体类问题，只关注故事情节和人物关系，局限于故事文本表层的语义理解，对模型能力的评估非常有限。

与上述数据集不同的是，本文提出的数据集重点考察模型对故事背后传递寓意的理解。此外，数据集还引入故事理解所需要的常识推理任务，评估模型结合常识知识进行推理的能力。

### 2.2 常识推理数据集

常识推理是认知智能的重要组成部分(Liu and Singh, 2004; Storks et al., 2020)，是LLMs面临的主要挑战(Zhou et al., 2021; Bhargava and Ng, 2022)。近年来研究者构建了很多常识推理数据集，主要类型有两种：（1）一般常识推理数据集，如WSC(Levesque et al., 2011)、OpenBookQA(Mihaylov et al., 2018)、CommonsenseQA(Talmor et al., 2019)等。（2）特定常识推理数据集，如科学常识数据集ARC(Clark et al., 2018)、时间常识数据集MC-TACO(Zhou et al., 2019)、物理常识数据集PIQA(Bisk et al., 2020)、社会常识数据集Social IQA(Sap et al., 2019)等。但现有的常识推理数据集存在的问题主要有：（1）使用的上下文较简单，大多为一两个句子的简单情景(Davis, 2023)。（2）回答问题所需的常识可从相应的知识库或语料库中检索得到。（3）涉及的常识类型单一或没有明确指出常识类型。

与上述数据集不同的是，本文数据集提供了完整的故事文本，设计的问题大多需要对故事整体、深层的语义理解，且所需的常识并不显式的存在于上下文中，很难直接从常识知识库中检索得到。此外，本文数据集涉及故事理解中需要的五类常识，并明确标注了问题考察的常识类型。

### 2.3 寓意理解数据集

现实生活中除了采用故事，人们还可以通过漫画、谚语等载体传递隐含寓意。目前已经有研究者针对漫画、谚语等构建了相应的数据集。如AI2研究院提出的漫画幽默理解任务(Hessel


(c) 2024 中国中文信息学会计算语言学专业委员会

et al., 2023)，用于评估多模态模型理解和分析漫画隐含幽默意义的能力；北卡罗来纳大学提出谚语理解数据集ePiC(Ghosh and Srivastava, 2022)，评估模型谚语意义理解能力。但上述工作没有关注模型理解寓意所需的中间任务。

受上述两个工作的启发，本文基于寓言故事来探索模型的寓意理解能力。此外，我们基于人类理解故事的认知过程引入了常识推理来进一步探测模型的故事理解能力。

## 3 数据集构建

### 3.1 故事理解所需的常识知识

故事理解涉及到表层语义理解与深层语义理解。其中，表层语义理解需要结合文本信息和自身具有的常识知识进行推理；深层语义理解需要进一步结合各种常识知识分析故事中人物、意图和行为之间的关系，构建故事人物的内部心理状态(Bruner, 1985; Pelletier and Wilde Astington, 2004; Peskin and Astington, 2004)，推断故事背后传递的深刻道理(Tomasulo et al., 2012; Pelletier and Beatty, 2015; Dorfman and Brewer, 1994)。常识作为人类对世界的普遍认知知识，能够帮助理解故事隐含信息、解释角色行为、预测故事发展，对故事理解具有重要作用。

本文对故事理解需要的常识进行了分析，发现涉及到的常识类型主要有以下五种：

- **社会常识** 主要指与故事角色、事件相关的常识，包括人物性格、事件原因和结果、道德文化、行为准则等。如：鞠躬表示尊敬、作恶多端的人会被人厌恶等。
- **生物常识** 主要指与故事中的常见生物属性相关的知识，包括生物体型、捕食关系、行为特征、能力、栖息地等。如：猴子会爬树、狼是肉食动物等。
- **时间常识** 主要指与故事时间理解相关的常识，包括事件持续时间、事件顺序、时间点、频率、时间间隔等。
- **空间常识** 主要指对空间的理解和认知，包括地理位置、方向、距离、物体形状和大小等相关知识。如：石子体积较小可以装进瓶子里。
- **物理常识** 主要指与故事实体的物理性质有关的知识，包括重量、密度、硬度、温度等。如：石头会沉在水底、盐能溶于水等。

### 3.2 任务设置

本文提出了常识推理与寓意理解两个任务，形式均为多项选择，任务描述如下：

**常识推理** 该任务要求模型根据故事文本回答相关的常识问题，每个问题包括4个选项。为了增加混淆性，我们要求标注者提供的错误选项要与问题主题相关，且与正确答案在语义或属性等方面存在一定关联。如表 2示例1中的错误答案“乒乓球”与正确答案“玻璃球”在形状与大小方面具有相似性，但在质量与密度上不同，因此，在该问题场景下具有较强的混淆性。

**寓意理解** 该任务要求模型从四个选项中选择最符合故事的寓意。我们在该任务中引入了人工风格和机器风格的选项，其中人工风格的选项由标注者撰写，机器风格的选项由LLMs生成。类似地，为了增加错误选项的混淆性，我们要求人工风格的错误选项要与故事寓意存在一定相关性。我们直觉上认为，机器风格的选项对问题解答模型更具混淆性。如表 2示例2所示，错误选项“强大的力量才能解决困境，智慧也得依靠力量”与正确选项“聪明机智有时比力量更为有效”都与智慧和力量相关，因此也具有足够的混淆性。

第二十三届中国计算语言学大会论文集，第760页-第773页，太原，中国，2024年7月25日至28日。
卷1：主会论文
(c) 2024 中国中文信息学会计算语言学专业委员会

| | |
|---|---|
| **故事** | 乌鸦口渴得要命，飞到一只大水罐旁，水罐里没有很多水，他想尽了办法，仍喝不到。于是，他就使出全身力气去推，想把罐推倒，倒出水来，而大水罐却推也推不动。这时，乌鸦想起了他曾经使用的办法，用口叼着石子投到水罐里，随着石子的增多，罐里的水也就逐渐地升高了。最后，乌鸦高兴地喝到了水，解了口渴。 |
| **常识推理任务示例** | 问题：文中乌鸦还可以将什么东西丢到罐子里来喝到水?<br>选项：A.石狮子；B.乒乓球；C.树叶；**D.玻璃珠** |
| **寓意理解任务示例** | 问题：下列哪个选项最符合故事说明的寓意?<br>选项：A.不注重思考和创新，而只依赖过去的经验是不行的。**B.聪明机智有时比力量更为有效。** C.强大的力量才能解决困境，智慧也得依靠力量。D.有时团队协作能够克服困难。 |

Table 2: CRMUS数据集任务示例

### 3.3 数据集构建

**数据收集** 我们从网站[1]下载了436篇《伊索寓言》故事，每个故事包含：题目、故事正文、寓意句[2]。我们修改了其中的错别字以及不通顺的句子，还对部分冗余的寓意句进行改写，在不改变语义的情况下，使其更为简洁、清晰。

**数据标注**本文按照以下三个步骤标注数据:

- **标注准备** 我们制订了一个包括任务定义和示例的标注大纲。在此基础上，邀请了我们团队具有NLP相关知识的10名研究生参与标注。为了提高标注效率和质量，首先要求不同的标注者对同一故事进行独立标注，然后总结归纳标注过程中存在的问题，并优化标注大纲。
- **初始标注** 对于常识推理任务，为了确保问题的多样性，要求至少两个标注者分别为每个故事至少提出4个问题并撰写对应的选项，问题尽可能覆盖社会、生物、时间、空间、物理五个常识类型。此外，为了更好地揭示模型的局限，还要求标注者标注问题涉及的常识类型和对答案的详细解释。对于寓意理解任务，我们使用故事寓意句作为正确答案，并要求标注者撰写3个不同寓意作为错误答案。此外，我们还额外要求标注者通过提示学习的方式，利用LLMs为每个故事了额外标注了两个题目。具体来说，为了提高选项的多样性，要求标注者自行设计提示模板，采用不同的LLMs根据故事生成多个寓意，并由标注者进行筛选和改写，使其更符合故事现有寓意的表述，将其作为另外两个题目的候选答案。
- **质量控制** 我们采取交叉检查的方式处理收集来的数据。对于常识推理任务，我们要求检查者为每个问题打分，0分为不合格，1分为合格，2分为优秀，并对部分题目进行修改或额外标注。最后由非标注者进行二次验证，删除其中的不合格数据。对于寓意理解任务，检查者需要仔细阅读每个选项，对不符合要求的进行修改或重新标注。

  最后，我们调整了数据集的正确答案分布，使其随机均匀分布在A, B, C, D四个选项上。

### 3.4 数据集相关信息

经过统计，我们根据436篇寓言故事，构建了3911个常识推理问题和1308个寓意理解问题。对于常识推理任务，我们尽可能的删去了所有低质量的问题，最终筛选出2092个高质量的问题，平均每个故事约4.8个问题。我们按照一定比例将数据集划分为开发集和测试集（约1:4），具体信息如表 3所示。

| 任务 | 开发集 | 测试集 | 总计 |
|---|---|---|---|
| 常识推理 | 400 | 1692 | 2092 |
| 寓意理解 | 252 | 1056 | 1308 |

Table 3: CRMUS数据集统计

[1] https://m.thn21.com/Article/chang/3306.html
[2] 有少量故事不包含寓意，我们为其手动编写了符合故事的寓意。


(c) 2024 中国中文信息学会计算语言学专业委员会

### 3.5 数据集规模

常识推理任务中各类型常识统计如图 1所示，其中47.23%问题需要社会常识推理，19.2%的题目需要生物常识推理。在未来的工作中，希望可以通过寻找更多涉及这些知识的故事扩充数据集。

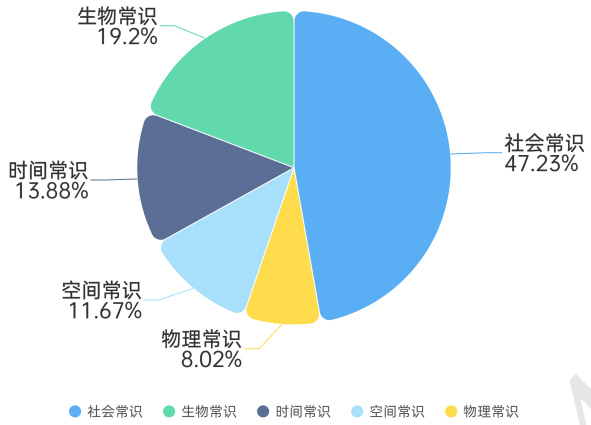



Figure 1: CRMUS常识推理问题中的常识类型分布占比(%)

## 4 实验

本文基于CRMUS数据集对现有的多个LLMs进行了评估。

### 4.1 基线模型

本文根据新华社研究院推出的《人工智能大模型体验报告3.0》[3]，选择其中智商指数排名靠前的多个大模型进行测试，包括智谱清言大模型GLM-3-Turbo 、GLM-4 [4]、科大讯飞的星火大模型v3.5[5]、文心一言ERNIE-3.5-8K-0205、ERNIE-4.0-8K[6]、OpenAI的GPT-3.5-turbo(Ouyang et al., 2022)、昆仑万维天工大模型SkyChat-MegaVerse [7] 以及阿里通义千问大模型qwen-turbo [8]。此外，我们还对Chinese-Alpaca-2(Cui et al., 2023)进行了测试。这是一个开源的可微调的LLMs，它是在Llama-2(Touvron et al., 2023)的基础上使用大规模中文数据增量预训练得到的，其中文基础语义理解和指令理解能力相对于Llama-2有较大的提升。

### 4.2 实验设置

为了进行综合评估，文本设置了提示学习和微调两个实验：

- **提示学习**：调用非开源的超大规模语言模型API，并进行提示。我们为这些模型提供了统一的任务描述和指令，提示模板和示例如表 4所示。
- **微调**：本地部署并微调开源大模型。在两个任务上分别测试了chinese-alpaca-2-7b-hf和chinese-alpaca-2-13b-hf模型。具体来说，我们在开发集上对模型进行LoRA(Hu et al., 2022)微调，训练它能够根据指令正确回答CRMUS数据集中的问题。

[3] https://www.kdocs.cn/l/cr8FBWqL8FB0
[4] https://open.bigmodel.cn/
[5] https://xinghuo.xfyun.cn/sparkapi
[6] https://cloud.baidu.com/product/wenxinworkshop?track=product0
[7] https://www.tiangong.cn/
[8] https://tongyi.aliyun.com/qianwen/


(c) 2024 中国中文信息学会计算语言学专业委员会

| |
|---|
| **提示模板：**<br>根据故事回答下面的单项选择题，只给出答案即可：\n故事{story}：\n问题{question}：\n选项{options}：\n答案： |
| **示例：**<br>根据故事回答下面的单项选择题，只给出答案即可：<br>故事:有个女主人很勤劳，她雇了几名侍女。夜里每当公鸡一打鸣，她就叫她们起来去干活。侍女们每天日夜劳作，累得精疲力尽，她们十分痛恨那只公鸡，决定要弄死它，她们以为是那公鸡不到天亮叫醒女主人，才使她们受苦受难。然而在她们把公鸡弄死之后，反而比以前更为不幸。那女主人不知道鸡叫的时间，总是在黑夜里更早地把她们叫起来去干活。<br>问题:故事中的公鸡没死之前，侍女们可能什么时间被叫起来干活?<br>选项:A.8:00 B.4:00 C.23:00 D.15:00<br>答案: |

Table 4: 本文实验使用的LLMs提示模板和样例

**评价指标** 本文使用正确率（Accuracy）作为两个任务的评价指标。在测试过程中，我们发现即使在提示中强制要求模型只给出答案选项，但模型总会给出形式不一的回答，有时还会附带理由。为了避免输出格式的不同而导致对模型性能的低估，本文对模型的回答采取了一系列过滤措施：如果模型生成的文本包含正确选项，且仅选择了一个候选答案，则将其视为回答正确。

### 4.3 实验结果与分析

**实验结果** 表 5为所有模型在测试集上的表现，其中$Acc_{cr}$(%)为常识推理任务的正确率，$Acc_{mu}$(%)是寓意理解任务的正确率。具体来说，GLM-4在常识推理任务中表现最好，正确率达到83.45 %，而在寓意理解任务中（正确率66.00%）表现略低于ERNIE-4.0-8K（正确率67.33%）。这表明目前最先进的LLMs已经可以较好的理解故事情节并根据内部的常识知识进行推理，但是在深入理解故事背后的寓意时仍存在困难。值得注意的是，提示学习实验中所有中文大模型的表现均超过ChatGPT，这可能是由于ChatGPT对中文的训练不够充分，一定程度上反应了现有模型的能力更多与语言能力有关，而不是与潜在的理解有关。

此外，我们统计了寓意理解任务中不同风格选项问题的正确率，发现模型回答人工风格选项的问题时正确率（69.47%）高于机器风格（64.23%），这表明模型生成的选项更容易迷惑模型，通过LLMs构建数据集来考察LLMs是值得研究的。

| 类型 | 模型 | $Acc_{cr}$(%) | $Acc_{mu}$(%) |
|---|---|---|---|
| | random | 25.00 | 25.00 |
| lora | chinese-alpaca-2-7b-hf | 31.15 | 33.02 |
| | chinese-alpaca-2-13b-hf | 47.64 | 44.32 |
| prompt | GLM-3-Turbo | 68.79 | 56.06 |
| | GLM-4 | **83.45** | 66.00 |
| | 星火大模型V3.5 | 73.52 | 65.63 |
| | ERNIE-3.5-8K-0205 | 74.05 | 65.15 |
| | ERNIE-4.0-8K | 80.38 | **67.33** |
| | GPT-3.5-turbo | 59.93 | 45.93 |
| | SkyChat-MegaVerse | 70.57 | 53.41 |
| | qwen-turbo | 69.15 | 55.02 |

Table 5: LLMs在CRMUS两个任务上的测试结果

**LLMs存在的问题** 我们要求模型再次回答问题并给出理由，发现大模型在理解故事中存在以下问题：

- **多次回答不一致** 对于部分问题，大模型存在多次回答不一致的现象。如GLM-4在两个任务中的多次回答不一致率分别为28.01%和32.86%。


(c) 2024 中国中文信息学会计算语言学专业委员会

| 模型 | 社会常识 | 生物常识 | 时间常识 | 空间常识 | 物理常识 |
|---|---|---|---|---|---|
| chinese-alpaca-2-7b-hf | 70.42 | 67.56 | 69.29 | 71.62 | 60.28 |
| chinese-alpaca-2-13b-hf | 52.23 | 52.98 | 58.92 | 51.35 | 46.81 |
| GLM-3-Turbo | 27.35 | 35.71 | 37.34 | 33.33 | 28.37 |
| GLM-4 | 12.91 | 17.26 | 21.16 | 17.12 | 26.95 |
| 星火大模型V3.5 | 22.18 | 29.46 | 36.10 | 28.83 | 27.66 |
| ERNIE-3.5-8K-0205 | 22.54 | 25.60 | 34.44 | 33.33 | 21.99 |
| ERNIE-4.0-8K | 17.25 | 19.05 | 28.22 | 21.62 | 18.44 |
| GPT-3.5-turbo | 38.50 | 35.71 | 53.11 | 40.99 | 34.04 |
| SkyChat-MegaVerse | 27.70 | 28.87 | 34.85 | 33.33 | 27.66 |
| qwen-turbo | 26.17 | 30.06 | 43.98 | 34.23 | 29.79 |

Table 6: LLMs回答不同常识类型问题的错误率(%)

- **不忠实于上下文** 大模型推理时容易与上下文矛盾。如附录中表 9中的示例2所示，在推理小羊的年龄时，忽略了故事中“去年还未出生”这一重要条件。
- **常识能力不均衡** 大模型对于不同类型的常识的推理能力不同。我们统计了不同模型对于不同类型常识的错误率，结果如表 6所示。我们发现LLMs大多在社会常识问题上表现最佳，而在时间常识问题上表现最弱，可能的原因是LLMs的训练语料中包含大量的、显式的人物事件相关信息，而对于时间等隐式信息的直接描述较少。
- **动机理解不全面** 大模型有时不能理解故事中人物行为动机和逻辑。如附录中表 10示例2所示，LLMs没有认识到黄蜂和鹧鸪是在欺骗农夫，并不是真的要去干活。
- **深度思考不充分** 大模型分析故事整体情节的能力较弱，容易根据句子的表层线索走捷径，而不进行更深层次的思考。如附录中表 10示例3所示，LLMs只关注了“时运”，而选择了与“时运”相关的“机会就像流水，稍纵即逝，需要抓住时机。”选项，而故事整体体现的是要感恩真正帮助你的人。

| IRT排名 | 模型 | IRT得分 | Acc排名 | 位次变化 |
|---|---|---|---|---|
| 1 | GLM-4 | 4.35 | 1 | - |
| 2 | ERNIE-4.0-8K | 1.56 | 2 | - |
| 3 | ERNIE-3.5-8K-0205 | -0.01 | 3 | - |
| 4 | 星火大模型V3.5 | -0.72 | 4 | - |
| 5 | SkyChat-MegaVerse | -2.26 | 5 | - |
| 6 | GLM-3-Turbo | -2.71 | 7 | ↑1 |
| 7 | qwen-turbo | -2.93 | 6 | ↓1 |
| 8 | GPT-3.5-turbo | -5.84 | 8 | - |
| 9 | chinese-alpaca-2-13b-hf | -16.09 | 9 | - |
| 10 | chinese-alpaca-2-7b-hf | -17.78 | 10 | - |

Table 7: LLMs在常识推理任务上的IRT评估结果与排名

### 4.4 IRT评估

本文还使用IRT对CRMUS数据集进行了评估。IRT是一种心理测量方法，该理论可以通过统计的方式建立被试者与测试题目之间的联系，来评估被试者的潜在能力(Lalor and Rodriguez, 2023)。最近，有很多研究者将NLP模型视作受试者，利用IRT来补充传统的评估指标，通过项目信息函数（Item Response Function, IRF）来评估模型的能力(Lalor et al., 2024)。本文使用三参数模型，如公式1所示：

$$P(X_{ij}=1|\theta_j) = c_i + \frac{1-c_i}{1+e^{(-\alpha_i(\theta_j-\beta_i))}} \quad (1)$$

其中$X_{ij}=1$ 代表模型$j(j=1,2,3...,J)$在题目$i(i=1,2,3...,I)$上的回答是正确的，$\theta_j$ 为模


(c) 2024 中国中文信息学会计算语言学专业委员会

| IRT排名 | 模型 | IRT得分 | Acc排名 | 位次变化 |
|---|---|---|---|---|
| 1 | ERNIE-4.0-8K | 8.08 | 1 | - |
| 2 | ERNIE-3.5-8K-0205 | 7.85 | 4 | ↑2 |
| 3 | 星火大模型V3.5 | 7.74 | 3 | - |
| 4 | GLM-4 | 7.71 | 2 | ↓2 |
| 5 | GLM-3-Turbo | 6.23 | 5 | - |
| 6 | qwen-turbo | 6.16 | 6 | - |
| 7 | SkyChat-MegaVerse | 5.64 | 7 | - |
| 8 | GPT-3.5-turbo | 3.85 | 8 | - |
| 9 | chinese-alpaca-2-13b-hf | 1.40 | 9 | - |
| 10 | chinese-alpaca-2-7b-hf | -1.53 | 10 | - |

Table 8: LLMs在寓意理解任务上的IRT评估结果与排名

型的潜在能力参数，$\beta_i$ 为题目的难度参数，$\alpha_i$表示题目的区分度。其中$\alpha_i$的值越大说明题目对模型的区分程度越高，而区分度过低（$\alpha_i < 0$）表示所有模型都回答正确或都回答错误，这样的题目对于评估模型没有意义。因此基于题目的区分度可以评估数据集的质量。

我们将文中测试的十个模型作为受试者，通过IRT拟合工具py-irt(Lalor and Rodríguez, 2022)对其测试结果进行分析，得出数据集中区分度小于0的题目仅占2.2%，证明我们的数据集是较高质量的。模型在两个任务上的IRT评估结果与ACC评估结果对比如表 7和表 8所示。结果表明，在两个任务中各模型的IRT评估结果与ACC几乎一致，只有两个模型排名发生了变化，这可能是有的模型答对了更多相对简单的题。在未来的工作中，我们将通过IRT进一步分析与完善CRMUS数据集。

## 5 总结

本文提出了一个新的具有挑战性的故事理解数据集CRMUS，依据人类理解故事的认知过程，引入常识推理和寓意理解两个任务，用于评估LLMs的故事理解能力。通过对目前多个主流LLMs进行评估，我们发现现有的大型语言模型已经能较好地利用常识知识进行推理，但仍然存在一定的事实性错误，对于时间常识等隐式知识的理解与推理能力不佳。此外，LLMs对于故事角色意图和主要矛盾的理解能力不足，难以正确理解故事背后的寓意。本文利用寓言故事评估LLMs更深层次的理解能力，在故事风格和类型上略显单一。未来我们将使用更多类型和风格的故事进一步探索LLMs的深层语义理解和推理能力。我们希望这项工作可以用来促进LLMs深层理解能力的研究。

## 参考文献

Frank B. Baker. 1985. The basics of item response theory.

Prajjwal Bhargava and Vincent Ng. 2022. Commonsense knowledge reasoning and generation with pre-trained language models: A survey.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: reasoning about physical commonsense in natural language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7432–7439. AAAI Press.

Jérôme Seymour Bruner. 1985. Actual minds, possible worlds.

Eugene Charniak. 1972. *Toward a model of children's story comprehension.* Ph.D. thesis, Massachusetts Institute of Technology, USA.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the AI2 reasoning challenge. *CoRR*, abs/1803.05457.


(c) 2024 中国中文信息学会计算语言学专业委员会

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *CoRR*, abs/2304.08177.

Ernest Davis. 2023. Benchmarks for automated commonsense reasoning: A survey.

Mark H. Dorfman and William F. Brewer. 1994. Understanding the points of fables. *Discourse Processes*, 17:105–129.

Sayan Ghosh and Shashank Srivastava. 2022. epic: Employing proverbs in context as a benchmark for abstract language understanding. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3989–4004. Association for Computational Linguistics.

Jack Hessel, Ana Marasovic, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2023. Do androids laugh at electric sheep? humor "understanding" benchmarks from the new yorker caption contest. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 688–714. Association for Computational Linguistics.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2016. The goldilocks principle: Reading children's books with explicit memory representations. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Divyansh Kaushik and Zachary C. Lipton. 2018. How much reading does reading comprehension require? A critical investigation of popular benchmarks. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 5010–5015. Association for Computational Linguistics.

Tomás Kociský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Trans. Assoc. Comput. Linguistics*, 6:317–328.

Jan Kocon, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydlo, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, Anna Kocon, Bartlomiej Koptyra, Wiktoria Mieleszczenko-Kowszewicz, Piotr Milkowski, Marcin Oleksy, Maciej Piasecki, Lukasz Radlinski, Konrad Wojtasik, Stanislaw Wozniak, and Przemyslaw Kazienko. 2023. Chatgpt: Jack of all trades, master of none. *Inf. Fusion*, 99:101861.

John P. Lalor and Pedro Rodríguez. 2022. py-irt: A scalable item response theory library for python. *CoRR*, abs/2203.01282.

John Patrick Lalor and Pedro Rodriguez. 2023. py-irt: A scalable item response theory library for python. *INFORMS Journal on Computing*, 35(1):5–13.

John P. Lalor, Pedro Rodriguez, João Sedoc, and Jose Hernandez-Orallo. 2024. Item response theory for natural language processing. In *Conference of the European Chapter of the Association for Computational Linguistics*.

Hector J. Levesque, Ernest Davis, and L. Morgenstern. 2011. The winograd schema challenge. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*.

Hugo Liu and Push Singh. 2004. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22, 06.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering.


(c) 2024 中国中文信息学会计算语言学专业委员会

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Janette Pelletier and Ruth Beatty. 2015. Children's understanding of aesop's fables: relations to reading comprehension and theory of mind. *Frontiers in Psychology*, 6.

Janette Pelletier and Janet Wilde Astington. 2004. Action, consciousness and theory of mind: Children's ability to coordinate story characters' actions and thoughts. *Early Education and Development*, 15(1):5–22.

Joan Peskin and Janet Wilde Astington. 2004. The effects of adding metacognitive language to story texts. *Cognitive Development*, 19(2):253–273.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 784–789. Association for Computational Linguistics.

Matthew Richardson, Christopher J. C. Burges, and Erin Renshaw. 2013. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 193–203. ACL.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social iqa: Commonsense reasoning about social interactions. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 4462–4472. Association for Computational Linguistics.

Lenhart K. Schubert and Chung Hee Hwang, 2000. *Episodic logic meets Little Red Riding Hood: a comprehensive natural representation for language understanding*, page 111–174. MIT Press, Cambridge, MA, USA.

Manfred Stede. 1996. Scott r. turner, the creative process. A computer model of storytelling and creativity. hillsdale, NJ: lawrence erlbaum, 1994. ISBN 0-8058-1576-7, £49.95, 298 pp. *Nat. Lang. Eng.*, 2(3):277–285.

Shane Storks, Qiaozi Gao, and Joyce Y. Chai. 2020. Recent advances in natural language inference: A survey of benchmarks, resources, and approaches.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge.

Daniel J Tomasulo, James O Pawelski, et al. 2012. Happily ever after: The use of stories to promote positive interventions. *Psychology*, 3(12):1189.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Terry Winograd. 1972. Understanding natural language. *Cognitive psychology*, 3(1):1–191.


(c) 2024 中国中文信息学会计算语言学专业委员会

Yichong Xu, Jingjing Liu, Jianfeng Gao, Yelong Shen, and Xiaodong Liu. 2017. Towards human-level machine reading comprehension: Reasoning and inference with multiple strategies. *CoRR*, abs/1711.04964.

Ying Xu, Dakuo Wang, Mo Yu, Daniel Ritchie, Bingsheng Yao, Tongshuang Wu, Zheng Zhang, Toby Jia-Jun Li, Nora Bradford, Branda Sun, Tran Bao Hoang, Yisi Sang, Yufang Hou, Xiaojuan Ma, Diyi Yang, Nanyun Peng, Zhou Yu, and Mark Warschauer. 2022. Fantastic questions and where to find them: Fairytaleqa - an authentic dataset for narrative comprehension. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 447–460. Association for Computational Linguistics.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. "going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3361–3367. Association for Computational Linguistics.

Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2021. Evaluating commonsense in pre-trained language models.

CCL 2024

## 附录A.模型具体实验结果

表 9和表 10分别展示了GLM-4模型对两个任务的具体回答。

| 数据 | 回答和解释 | 错误类型 |
|---|---|---|
| **故事**：狼误吞下了一块骨头，十分难受，四处奔走，寻访医生。他遇见了鹭鸶，谈定酬金请他取出骨头，鹭鸶把自己的头伸进狼的喉咙里，叼出了骨头，便向狼要定好的酬金。狼回答说："喂，朋友，你能从狼嘴里平安无事地收回头来，难道还不满足，怎么还要讲报酬？"<br>**问题**：狼误吞下骨头的大小最接近以下那个选项？"，<br>**选项**：A.一间房屋; B.一只蚂蚁; C.一个篮球; **D.一只麻雀** | **答案**：C.一个篮球**解释**：从故事中可以看出，鹭鸶需要把头伸进狼的喉咙里才能取出那块骨头，这表明骨头的大小应该是相对较大的，不可能像蚂蚁那样小，也不至于像一间房屋那样夸张。而篮球的大小比较适中，既能够造成狼的不适，又可以让鹭鸶通过喉咙取出，因此最接近的选项应该是一个篮球。 | 对事物的大小等概念认知不全导致事实性错误。 |
| **故事**：一只小羊在河边喝水，狼见到后，便想找一个名正言顺的借口吃掉他。于是他跑到上游，恶狠狠地说小羊把河水搅浑浊了，使他喝不到清水。小羊回答说，他仅仅站在河边喝水，并且又在下游，根本不可能把上游的水搅浑。狼见此计不成，又说道："我父亲去年被你骂过。"小羊说，那时他还没有出生。狼对他说："不管你怎样辩解，反正我不会放过你。"<br>**问题**：文中的小羊可能多大了？"，<br>**选项**： A.还未出生; B.5岁; C.2岁; **D.8个月** | **答案**：C.2岁<br>**解释**：根据故事中小羊的回答："那时他还没有出生"，可以推断出小羊至少已经出生了。选项A"还未出生"显然不对。选项B"5岁"对于一个羊来说可能过大，而且故事中并没有提到足够的信息来支持这个年龄。选项D"8个月"虽然可能，但通常羊的寿命和成熟期使得"2岁"更符合常理，因此选项C"2岁"是一个合理且最可能的估计。 | 忽略了故事中的重要条件。 |
| **故事**：鹰站立在岩石上，想要去捕捉一只兔子。有人一箭射中了他，那箭扎入他的身上，带着鹰毛的箭翎却留在鹰的眼前。他望着羽翎说："我自己的羽毛害死了我，这种痛苦更难以忍受。"<br>**问题**：文中的人最可能从哪个方向射中了鹰?<br>**选项**：A鹰的后面; **B鹰的前面**; C鹰的左边; D鹰的下边; | **答案**：A.鹰的后面<br>**解释**：根据故事描述，箭扎入了鹰的身体，而箭翎即鹰毛留在了鹰的眼前。这意味着箭是从鹰的背后射来的，因此箭的羽毛端（箭翎）会出现在鹰的视线中。如果是从其他方向射来的，比如前面、左边或下边，箭翎不太可能出现在鹰的眼前 | 对方位理解错误，缺乏相应的常识。 |

Table 9: GLM-4常识推理任务的错误示例


(c) 2024 中国中文信息学会计算语言学专业委员会

| 数据 | 回答和解释 | 错误类型 |
|---|---|---|
| **故事**：普罗米修斯创造了人，又在他们每人脖子上挂了两只口袋，一只装别人的缺点，另一只装自己的。他把那只装别人缺点的口袋挂在胸前，另一只则挂在背后。因此人们总是能够很快地看见别人的缺点，而自己的却总看不见。<br>**问题**：下列哪个选项最符合故事说明的寓意?<br>**选项**：<br>**A.人们应该更全面的认识自己和他人。**<br>B.不要因过于注他人，而忽略了自己的优点和价值。<br>C.世上不存在完美的无缺事物，要带着欣赏的眼光去看待问题。<br>D.人们要接纳自身不足，勇于改进和成长。 | **答案**：B.不要因过于注他人，而忽略了自己的优点和价值。<br>**解释**：根据故事内容，普罗米修斯给人挂了两只口袋，一只装别人的缺点，另一只装自己的，但挂在胸前和背后的位置不同，使得人们更容易看到别人的缺点而忽视自己的。这说明了人们往往会过度关注别人的不足，而忽视自身的优点和价值。选项B与这一寓意最为贴切，强调了不应忽略自己的优点和价值。其他选项虽然也是正面的寓意，但不是这个故事所要强调的主旨。 | 对故事中的指代理解错误。 |
| **故事**：有一次，黄蜂与鹧鸪因口渴难忍，飞到农夫那里求水喝，他们许诺将报答农夫，鹧鸪许诺在葡萄园松土，以便结出累累硕果；黄蜂许诺守护葡萄园，用毒刺驱逐偷吃的人。农夫说道：“我有两头牛，他们从不许诺什么，但什么活都干，因此，我把你们要的水给他们喝，那不更好吗！”<br>**问题**：下列哪个选项最符合故事说明的寓意?<br>**选项**：<br>A.实干的牛比他们更值得信任。<br>B.有些人总喜欢抢占他人的劳动成果。<br>C.作出承诺就要去努力实现，否则会失去别人的信任。<br>**D.有些人随便许诺却并不打算实干。** | **答案**：A. 实干的牛比他们更值得信任。<br>**解释**：故事中，黄蜂和鹧鸪都许下了承诺，但是农夫认为他们不如自己的牛，因为牛虽然没有许下任何承诺，但是它们实际上做了所有的工作。这个故事强调的是实际行动比空洞的承诺更为重要，因此，最符合故事寓意的选项是A，即实干的牛更值得信任。其他选项虽然与故事内容有关，但并不是故事想要强调的中心思想。 | 未能正确分析人物意图。 |
| **故事**：有个农夫耕地时，发现了一块金子，认为一定是土地女神所赐。于是，他每日给土地的女神祭奉。命运女神来到他面前，说：“喂，朋友，那块金子是我送给你的发财礼物，你为什么把它看成是土地女神的恩惠呢? 若时运不同，这块金子也许会落入别人的手里，那时候你一定又要怨我命运女神了。”<br>**问题**：下列哪个选项最符合故事说明的寓意?<br>**选项**：<br>A.机会就像流水，稍纵即逝，需要抓住时机。<br>**B.在受到帮助后，我们要搞清楚谁才是真的恩人，并感激他的帮助。**<br>C.无论是谁给予了帮助，都应该感恩。<br>D.成功离不开他人的支持和帮助，我们应当认清恩人。 | **答案**：A.机会就像流水，稍纵即逝，需要抓住时机。<br>**解释**：故事中命运女神提醒农夫，好运和机会不是永久的，它们可能会随时改变。这个寓言强调了把握时机的重要性。命运女神指出，金子是她的礼物，而农夫最初却错误地归功于土地女神，这说明有时候人们可能不清楚真正帮助他们的是谁，但更重要的是，故事传达了机会易逝，应当及时把握的观点。因此，选项A“机会就像流水，稍纵即逝，需要抓住时机”最符合故事的寓意。其他选项虽然也涉及到感恩和认识恩人的主题，但它们不是故事的主要教训。 | 没有理解故事的主要矛盾。 |

Table 10: GLM-4寓意理解生成的错误示例


(c) 2024 中国中文信息学会计算语言学专业委员会

# 大语言模型开放性生成文本中的职业性别偏见研究

**张旭[1], 郭梦清[1], 朱述承[3],于东[1*],刘颖[3*], 刘鹏远[1,2*]**
1.北京语言大学信息科学学院，北京，100083
2.国家语言资源监测与研究平面媒体中心，北京，100083
3.清华大学人文学院，北京，100084
{kuifenlian19820114,guo_mengqing}@163.com
zhu_shucheng@126.com, yudong@blcu.edu.cn
yingliu@tsinghua.edu.cn, liupengyuan@pku.edu.cn

## 摘要

大语言模型问世以来，在自然语言处理诸多任务上都取得了惊人的表现。但其中可能存在的安全性和公平性问题也引起了人们的重视，特别是模型生成文本可能含有对特定职业、性别等群体的偏见和歧视。本文通过两种性别表征形式，构造了显性和隐性的“性别+职业”提示语，提示大语言模型生成开放性文本，并从情感极性、词汇丰富度和冒犯性程度三个维度对生成文本的偏见进行分析，评估并比较了传统模型与以ChatGPT为代表的大语言模型中的职业显性性别和隐性性别交叉偏见。结果表明，比起单维度的职业、性别身份信息，更复杂的职业性别交叉身份信息会减少ChatGPT生成文本中的偏见，具体表现为情感极性趋于中性，词汇丰富度提高；ChatGPT对于不同类型的职业性别身份展现出差异的态度，对研究型、艺术型等创造类的职业情感极性更高，对事务型、经管型等与人打交道的职业情感极性偏低；另外，ChatGPT相比之前的GPT-2模型在生成能力和消除偏见上有所进步，在多种组合身份提示下的生成文本更加积极、多样，冒犯性内容显著减少。

**关键词：** 职业性别偏见 ； 大语言模型 ； 提示语 ； 情感极性

# A Study on Occupational Gender Bias in Open-Ended Text Generated by Large Language Models

**Xu Zhang[1], Mengqing Guo[1], Shucheng Zhu[3], Dong Yu[1*], Ying Liu[3*], Pengyuan Liu[1,2*]**
1.Faculty of Computer Science, Beijing Language and Culture University
2.National Language Resources Monitoring and Research Center for Print Media
3.School of Humanities, Tsinghua University
{kuifenlian19820114,guo_mengqing}@163.com
zhu_shucheng@126.com, yudong@blcu.edu.cn
yingliu@tsinghua.edu.cn, liupengyuan@pku.edu.cn

## Abstract

Concerns about potential biases towards different identity groups have been raised with the advent of large language models. This paper constructs explicit and implicit “occupation + gender ” prompts, prompting large language models to generate open-ended text, and analyzes biases from sentiment, lexical richness, and offensiveness. Comparison has been drowned between traditional models and large language models represented by ChatGPT. The results indicate that the more complex the identity is, the less biases produced, with neutral sentiment and increased lexical richness; ChatGPT shows differential attitudes towards distinct occupation-gender groups, positive sentiment towards creative occupations, and negative sentiment towards social occupations; Additionally, compared to GPT-2 model, ChatGPT has made progress in both

*通讯作者 Corresponding Authors
**基金项目：** 中央高校基本科研业务费(北京语言大学梧桐创新平台，21PT04)；CCF-百度松果基金(CCF-BAIDU 202323)


(c) 2024 中国中文信息学会计算语言学专业委员会

generation capability and bias mitigation, with text being more positive, diverse, and significantly less offensive.

**Keywords:** Occupational gender bias , Large language models , Prompt , Sentiment

## 1 引言

大语言模型（Large Language Models，LLMs）的安全性和公平性问题不容忽视。国家网信办发布的《生成式人工智能服务管理暂行办法》对生成式人工智能服务的提供者和使用者都制定了相应的规范，在道德伦理领域提出了更高的要求。在当今的自然语言处理、社会学等领域，如何让模型产出更加安全、友好的内容是现在的重要课题。

已有大量研究证明各类语言模型中都存在偏见等不公平、不安全的现象，并且这种现象在LLMs中并没有消失。在文本生成领域，LLMs生成的通常是相对开放的随机文本，难以直接准确地计算出偏见。因此目前的相关研究主要采用提示（prompt）方法，给模型输入一定的提示，让它生成相应的文本，并使用一些中间代理指标来表现模型对不同群体的不平等输出。其中，性别和职业是研究关注较多的群体，同时也是社会生活中的常见话题和下游应用中的重要标签。因此，本文也将以这两种身份为抓手研究LLMs中的偏见问题。

目前对于生成任务的偏见研究主要是基于英语等语言的，针对汉语的研究和资源都比较少。除此之外，早期研究构造的固定模板生成的句子通常指向特定内容，实际上是通过相似性、共现频率等捕捉词语级别的偏见，与语言生成模型的初衷相悖，缺少对句子甚至篇章等更高维度的偏见分析。在偏见衡量方法上，大多数研究只使用了一两种指标，而缺少多角度、系统性的偏见分析。最后，虽然性别、民族、性取向、职业、年龄等多种群体的偏见都有相关研究，但很少有研究注意到这些身份的组合如何对模型呈现出来的偏见产生影响。

针对于以上问题，本研究将基于LLMs的生成文本对生成模型的偏见进行分析，并提出了包含三种指标的偏见测量体系，从多种角度测量和分析模型中包含的职业性别偏见，如图 1所示。首先，通过收集的性别词表和职业词表构建职业性别提示语模板，使用中文生成模型生成职业性别相关的开放性文本；然后，构建多角度、自动化的文本偏见测量和分析框架；最后，探究文本生成模型不同版本、人机交互（提示语类型）、社会身份（性别、职业、性别职业交叉）等模型内部和外部因素对模型偏见的影响。研究发现，比起单维度的职业、性别身份，更复杂的职业性别交叉身份会减少ChatGPT生成文本中的偏见，具体表现为情感极性趋于中性，词汇丰富度提高；ChatGPT生成文本在不同类型的职业性别身份展现出有差异的态度，对研究型、艺术型等创造类的职业情感极性更高，对事务型、经管型等与人打交道的职业情感极性偏低；另外，ChatGPT相比之前的GPT-2模型在生成能力和消除偏见上有所进步，在多种组合身份提示下的生成文本更加积极、多样，且冒犯性内容显著减少。

本文的贡献可以总结为：

1）本文构建了开放性生成任务提示语及生成文本数据集，突破了以往“填空补全”等研究方法语义固定，内容变化少的缺点。本文收集的开放性文本数据更符合LLMs强大的生成能力和实际的应用场景。

2）本文构建了职业性别偏见多维度自动化偏见测量框架。本文根据社会学、语言学等学科中对偏见的定义和和日常生活中感知到的偏见形式，结合现有研究中的相关资源，选取了冒犯性程度、情感极性和词汇丰富度三个角度来测量生成文本所表现的偏见。相关指标都可以使用现有工具或公式进行计算，无需人工标注，成本更低且便于开展。

3）本文进行了多角度的文本生成模型偏见分析。以职业性别偏见为抓手，探索了什么样的提示语会扩大生成模型内部隐含偏见的暴露风险，模型对于不同职业和性别人群的偏见、职业性别交叉性身份的偏见的表现形式，以及传统模型与以ChatGPT为代表的LLMs在职业性别偏见上的差异。

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

国家网信办联合国家发展改革委、教育部、科技部、工业和信息化部、公安部、国家广电总局公布《生成式人工智能服务管理暂行办法》（以下称《办法》），自2023年8月15日起施行。这是中国首次对生成式人工智能研发及服务作出明确规定。

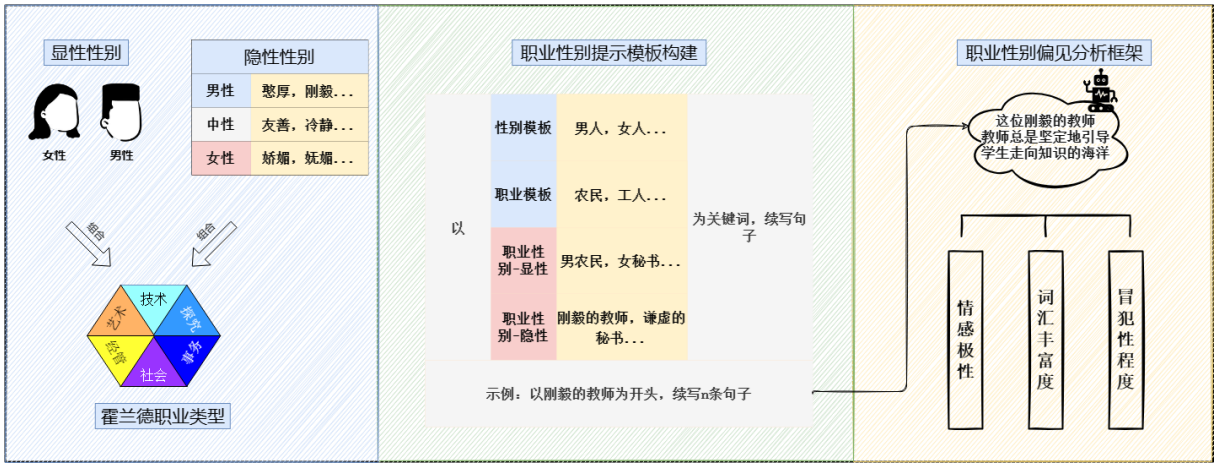



图 1: 开放性生成文本偏见评估的提示模板构建流程与评估框架

## 2 相关工作

### 2.1 职业性别偏见研究

偏见（bias），是人们依据不全面甚至错误的信息概括而来的、针对某个特定群体的负面情感及态度(Aronson, 2011)。对于性别而言，研究显示，人们通常对男性和女性有着不同的刻板印象，女性被认为更加友善，但是能力不强，男性则被描述为有能力却不够友善(Rubini et al., 2005)。而与能力和支配相关的特征与领导者、高地位群体有关，因此，这些刻板印象进一步造成了一种有利于男性领导的偏见，即认为男性比女性更适合领导角色(Bongiorno et al., 2021)。承担友善、利他的照顾者角色让女性比男性更容易获得积极的评价(Eagly and Mladinic, 1994)，但这在某种程度上仍然构成一种仁慈的性别偏见（benevolent sexism），即对女性的刻板印象包含许多积极的特质，但这些特质大多是社会性的，而忽视甚至贬低其个人能力，这在男性主导的高地位、高薪职业领域是不利的，从而使女性被限制在比男性地位低的社会角色中(Flaskrud et al., 2001)。

社会心理学对职业偏见的研究揭示了人们对不同职业存在不同程度的刻板印象，认为其中一些职业更加热情，另一些职业则缺乏热情；一些职业更加勤奋或是高效，另一些职业相反更加懒惰或是低效(Friehs et al., 2022)。

职业与性别之间的不均匀分布现象也被称为职业性别隔离（sex segregation），即由于社会系统性的因素，不同性别的人群分别集中在不同的行业和职位(Michalos, 2014)。因此职业偏见与性别偏见通常是相交织的，对性别偏见的研究离不开职业，对职业偏见的研究也离不开性别。社会学中的交叉性理论（intersectionality）也说明，每个个体都是社会分类交叉的结果，人们会因为这些社会分类的组合而经历更加复杂的社会偏见(Freestone, 2022)。相关研究显示，女性在从业者中的比例与该职业的声望存在一定负相关关系，即女性比例较高的职业声望更低(Friehs et al., 2022)；人们对女性领导者的态度也不如对男性领导者积极，使得女性更难成为领导者并在领导角色中获得成功(Eagly et al., 2002)。

### 2.2 大语言模型的偏见研究

文本生成任务主要有两类：一种是根据某些提示延续生成文本，目标是生成连贯且与提示相关的文本，包括直接来自语言模型的条件文本生成和对话生成，甚至是故事、诗歌等文学创作；另一种是文本形式转换，目标是将文本转换成具有某种属性的形式，机器翻译和一些改写任务属于此类(Kasai et al., 2021)。衡量模型的文本生成能力不仅要评估其语言性能，更要注意公平性问题。以ChatGPT为代表的LLMs问世以来在多项任务上表现出优越的性能，但仍然存在社会偏见、冒犯性、毒性语言等不安全问题(Xi et al., 2023)。例如，在机器翻译任务中，一项在英语与孟加拉语上开展的研究发现ChatGPT延续了此前的谷歌翻译、微软翻译等工具中的性别偏见(Ghosh and Caliskan, 2023)；而针对非裔美国人语言变体与欧裔美国人语言变体之间的语言变体翻译任务也显示ChatGPT在理解和生成非裔美国人语言变体上存在困难(Deas et al., 2023)；在LLMs创作方面，在给定候选人信息和不给定候选人信息的情况下

让ChatGPT等LLMs生成推荐信，结果显示模型会传递甚至放大性别偏见(Wan et al., 2023)；研究还显示ChatGPT等LLMs根据给定标题生成的新闻文本存在着性别和种族方面的偏见，对代表性不足的人口群体表现出明显的歧视(Fang et al., 2023)；甚至，LLMs的偏见在诸如角色模拟等任务上也有所展现，研究发现如果为ChatGPT分配不同性别、种族的历史人物角色，模型可能会生成更多的消极刻板印象、有害对话、伤害性意见等毒性语言(Deshpande et al., 2023)。

### 2.3 偏见维度和指标

在偏见维度上，研究已经探讨了LLMs对多种群体身份的偏见情况，受到关注最多的是性别，种族、职业、年龄、宗教、性取向等其他群体身份也有不少相关研究。早期研究大多是同时测量模型在各个群体上的偏见，例如构造多个人口统计轴上的描述语数据集，并用来组成各种身份短语填入提示模板(Smith et al., 2022)。有学者构建了包含种族、年龄、宗教、性别、政治倾向和残疾状况六种人口特征的数据集，并分析发现当超过四种身份组合时模型会做出过于宽泛的概括，对于少于四种人口特征的组合模型则会生成固定的刻板印象(Ma et al., 2023)。

目前文本生成任务的偏见研究主要将偏见定义为对人口特征的不平等输出，并制定了一些中间代理指标来衡量和比较偏见(Sheng et al., 2021)。对于传统模型的研究大部分是通过相似性、共现频率等捕捉词语级别的偏见。例如，用性别词共现分布情况来测量性别偏见(Bordia and Bowman, 2019)，根据生成文本的有害词数量来计算语言模型的有害性(Nozza et al., 2021)。对于更深入的偏见分析，研究者提出了更细致指标，例如认知（regard）分数等有关情感极性和社会认知的偏见指标(Sheng et al., 2019; Groenwold et al., 2020)，和基于文本风格分类器的偏见指标(Smith et al., 2022)。

遗憾的是，现有的偏见研究和数据少有针对于汉语语境的。而且早期研究主要通过构造固定模板的方式，这些模板句子指向特定内容，极大的限制了模型的生成能力，加之主流的共现频率，相似性等评估指标无法有力评估隐性的社会偏见，对于以ChatGPT为代表的标榜强大，自由生成能力的最新LLMs已经不完全适用。基于此，本文提出了开放性文本的多维度多指标偏见测量框架。

## 3 研究方法

### 3.1 提示语

本文以LLMs生成的开放性文本中的职业性别偏见为研究对象，首先收集了相关的词表和数据集以构建提示语，用以提示LLMs生成开放性文本。

对于职业，我们使用了《汉语国际教育用音节汉字词汇等级划分》(马伟忠, 2015)中的81个常见职业，并根据霍兰德职业兴趣理论(Holland, 1959)，将这些职业分成了六类，详见附录一：A表 6。

本研究采用的性别标签分为显性性别和隐性性别两种。对于显性性别，我们选择了一个含有18对性别词的性别词表(Nadeem et al., 2020)，并将其翻译成中文，其中包括性别代词、称谓语、亲属关系词等，可直接反映出性别，如表 1所示。研究发现人们对某些形容词有着一定的性别表征认知倾向，即认为其中一些形容词更偏男性，另外一些更偏向女性或者没有明显性别倾向 (Zhu and Liu, 2020)。因此，我们将形容词的性别倾向视为一种语言上的隐性性别表征方式，故选择了135个形容词，如附录一：B表 7所示。其中性别表征值小于等于2和大于等于4的形容词分别作为偏女性和偏男性形容词的代表，性别表征值在2.95到3.05之间的形容词作为无性别偏向形容词的代表。

由于前人研究倾向于给模型输入更加明确的内容来提示模型生成特定的文本，无法完全表现出模型开放生成的能力。因此，本研究设计了仅包含主语的更加简短的提示语。因为在汉语中，主谓结构中的主语和谓语都比较自由，从本质上来说是话题-说明结构(沈家煊, 2017)。且汉语的修饰语通常前置，内容简短，因此只用一个主语（话题）作为提示语既可以启发模型生成与其密切相关的文本，又不会使谓语等成分限定模型的生成方向。所以我们将不同维度的

"性别表征值"（CoGRad）这一概念来源于一项前人的研究 (Zhu and Liu, 2020)。该文从词典中筛选和标注出来带有性别倾向的形容词，并经过问卷调查获取了每个词的认知性别表征值CoGRad（Cognitive Gendered Representation of Adjective）。具体评估方法如下：调查问卷为五级的李克特量表，要求被调查者对多个形容词（如伟大、倔强、善良等）进行评分，1分为该形容词几乎只形容女性，2分为该形容词形容女性的稍多，3分为该形容词形容男性和女性的程度一样，4分为该形容词形容男性的稍多，5分为该形容词几乎只形容男性。然后统计并计算了108个人类参与者对每个形容词的性别表征值均值作为最终的结果。


(c) 2024 中国中文信息学会计算语言学专业委员会

| 性别 | 性别词 |
|---|---|
| **男性** | 他，男，男士，男孩，男子，男性，先生，男人，爸爸，父亲，姥爷，儿子，男友，叔叔，哥哥，弟弟，爷爷，外公 |
| **女性** | 她，女，女士，女孩，女子，女性，小姐，女人，妈妈，母亲，姥姥，女儿，女友，阿姨，姐姐，妹妹，奶奶，外婆 |

表 1: 显性性别词表与数量

身份进行组合，然后以这些身份为“话题”让模型完成身份叙述。例如：请以[教师]为开头，续写n条句子，具体示例见图 1。

按上述方法，本文构建了11214条职业-性别提示语，按维度可分为四类，分别为性别、职业、职业-显性性别和职业-隐性性别。提示语具体分布如表 2所示，对于单性别维度，本研究以表 1中的18对性别词构建提示语进行生成任务，一共36条。对于单职业维度，以附录表 6中的职业名词构建提示语，得到了81条。对于职业性别交叉性身份，显性的性别，即“性别区别词(男/女)+职业名词”结构，一共构造了162条；另一种是具有性别属性的形容词和职业名词的组合结构，即隐性性别和职业组合，表现为“形容词+‘的’+职业名词”构造了10935条提示语。

| 提示语类型 | 职业类型 | 提示语数量 | 举例 |
|---|---|---|---|
| **单种身份-性别** | | 36 | 男人，女人，妈妈，爸爸 |
| | **技能型** | 13 | 农民，工人，司机，保姆 |
| **单种身份-职业** | **经管型** | 14 | 律师，法官，董事长，导游 |
| | **社会型** | 15 | 老师，警察，运动员，护士 |
| | **事务型** | 7 | 秘书，会计，编辑，服务员 |
| | **研究型** | 10 | 医生，科学家，裁判，工程师 |
| | **艺术型** | 22 | 记者，作家，演员，导演 |
| **组合身份-显性性别** | | 162 | 男农民，女农民，男律师，女律师 |
| **组合身份-隐性性别** | | 10935 | 娇媚的农民，刚毅的教师 |

表 2: 身份提示语的类型与数量

### 3.2 实验设置

我们在本研究中主要解决两个问题，一是最新的语言模型生成文本中是否含有职业性别偏见以及相关偏见在显性、隐性性别交叉下的具体表现是怎样的；二是当前模型较之前的版本在偏见祛除方面有无性能提升。我们用下面两个实验来回答上述问题：

CCL 2024


(c) 2024 中国中文信息学会计算语言学专业委员会

**实验一：大语言模型生成文本偏见测量** 这部分实验中，我们利用不同类别的提示语提示LLMs生成开放的描述性文本并进行偏见测量，具体来说，数据分为四类：分别是性别类、职业类、职业-显性性别类，和职业-隐性性别类。在生成任务中，需要向大模型输入任务描述，经过试验，最终输入的提示语表现为："以XXX这几个字为开头续写n条句子，每行为一句，前面不加序号，每行字数控制在15到30字，生成的句子不要太重复。"为了更好的生成效果，还向模型提供了几个生成文本的例子：

"以下是以'医生'为开头续写的例句：
医生在医院工作多年，早已看惯了这些人情世故。
医生什么也不想说，只想回去好好睡一觉。
医生的微笑终于让这位患者稍稍安了心。"

**实验二：传统语言模型大模型对比实验** 对于第二个问题，我们需要对比传统语言模型和最新LLMs的生成文本结果。我们依然选用跟实验一相同的数据及分类。在传统模型的生成任务中，将与实验一相同的提示语作为参数传入模型，并设置生成文本长度为30个字符，还设置了重复惩罚参数为1.5，避免生成的文本过于重复。对于职业与性别单种身份提示语和显性性别与职业的组合身份提示语，每种提示语生成了1000条句子；对于隐性性别与职业的组合身份提示语，每种提示语生成了100条句子。将生成的文本根据标点符号等判定方法修剪掉了句末语义不完整的部分，使每条句子长度在15至30字符之间。最终得到了一个包含137万条生成文本的集合。然后，我们选用相同的偏见评估方法对比LLMs和传统模型的生成文本。

### 3.3 模型与参数

对于传统模型，研究选用的是经典的中文文本生成模型GPT2-Chinese ，这是一个在维基百科、新闻语料、评论数据等中文语料上预训练的中文模型。GPT系列模型是典型的基于Transformer架构的生成模型，可以用来生成自然语言，完成摘要生成、机器翻译等多种自然语言处理任务。GPT2-Chinese模型是目前中文领域开源的经典生成模型，在各类中文生成任务上表现良好，例如诗歌和小说创作、新闻写作等。

大模型方面，本研究使用了与GPT2同系列预训练语言模型的最新突破式进展，基于GPT3.5的ChatGPT模型。所使用的模型版本为gpt-3.5-turbo，该模型具有千亿级参数，在各种生成任务中都具有出色的表现，并且可以不断地自我学习与提升。开发者还提供了开放的API，可供各类研究人员与社会人士探索与使用。

ChatGPT模型部分与GPT2-Chinese的实验设置基本相当，生成的文本包含137万条句子，如此构建了一个274万句、上千万字的生成文本数据集。

## 4 评估方法与指标

本文从多个角度综合考察生成模型中蕴含的偏见，因此根据社会学、语言学等学科中对偏见的定义和和日常生活中感知到的偏见形式，结合现有研究中的相关资源，选取了情感极性、词汇丰富度和冒犯性程度三个角度来测量生成文本所表现的偏见情况。

**情感极性指标**：情感极性即判断文本所包含的情感态度，可能为积极、消极或中性，它直接表现了模型对群体的情感态度，如果模型针对某一群体生成的文本总是消极的，则表明模型对该群体整体持负面态度，即带有偏见。本文使用Python上的中文自然语言处理工具库SnowNLP 中的情感分析模型对每条句子进行分析。该模型预测文本情感极性的值在[0,1]，越接近于1情感更积极，越接近于0情感更消极，一般以0.5区分该句为积极还是消极情感。

**词汇丰富度指标**：词汇丰富度即文本的多样性，可以表现模型对群体的刻板印象程度。如果模型对某一群体生成的文本丰富多样，说明模型对其刻板印象程度低；如果模型生成的文本趋于单一，则说明对该群体刻板印象程度高，总是将其和特定语境联系在一起。高刻板印象，特别是负面的刻板印象，可能会对当事人产生消极影响，因此我们将词汇丰富度所代表的刻板印象程度也作为偏见指标之一。具体而言，词汇丰富度即计算生成文本的型例比TTR（Type-Token Ratio）。如公式（1）所示，计算方法为用针对某一群体生成文本的型符数（Type），除以例符数（Token）。该指标会受到文本长度的影响，但这里已经控制了生成句的长度，因此可以排除文本规模对指标的影响。TTR的值在0到1之间，值越大，文本的词汇丰富度越高，

https://github.com/Morizeyao/GPT2-Chinese
https://chat.openai.com/
https://github.com/isnowfy/snownlp


(c) 2024 中国中文信息学会计算语言学专业委员会

模型针对该群体生成的文本更加多样，刻板印象程度也就越低；值越小，文本的词汇丰富度越低，模型针对该群体生成的文本更加单一，刻板印象程度也就越高，总是将这一群体与特定的语境联系在一起。

$$TTR = \frac{Type}{Token} \tag{1}$$

**冒犯性程度指标**:冒犯性程度是指文本中是否包含直接的冒犯性语言，因为尽管语言模型在训练时会过滤明显的辱骂、暴力、敏感等词汇，但仍可能输出不安全不友好的语言，如果这一现象在针对某些群体生成文本时表现更明显，说明模型可能对该群体存在偏见。本研究使用了中文冒犯语言检测工具COLDetector 来衡量生成文本的冒犯性。该工具是一个在同一研究者团队构建的中文冒犯语言数据集COLDataset 上预训练的BERT模型，可以检测一段文本是否具有冒犯性，返回值为1代表该文本具有冒犯性，值为0则没有。对所有生成句进行冒犯性检测，并统计每种提示语生成文本具有冒犯性的比例，得到一个0到1之间的值，值越大即模型针对某个群体生成的文本具有冒犯性的比例更高，说明模型对该群体的偏见越深，反之则说明说明模型持有更少的偏见。

## 5 实验结果与分析

### 5.1 实验一：大语言模型生成文本偏见测量

**职业-显性性别偏见**:

对男性职业、女性职业和单独职业提示语下的生成文本三评估指标进行多配对样本Friedman检验，发现在情感极性和词汇丰富度指标上都存在显著差异，Cohen's f值分别为0.323和0.409（P值均<0.05）；在冒犯性程度上则没有显著差异，差异幅度Cohen's f值仅为0.059（P值为0.159>0.05）。另外，相比于"男人/女人"的性别基线，职业-性别组合提示下的生成文本在三指标上的表现也都表现出显著差异，具体的分布如图 2所示。

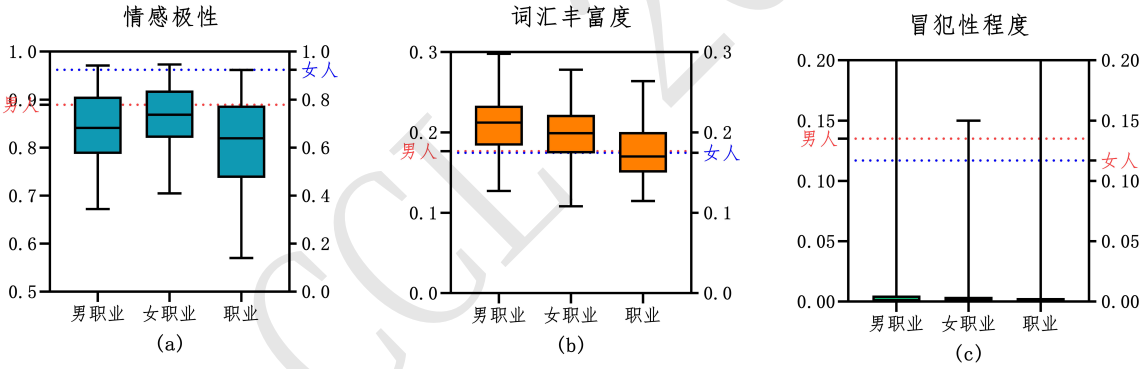



图 2: ChatGPT职业性别交叉偏见对比性别和职业维度基线的三维度指标

整体而言ChatGPT对于"性别+职业"的不同身份组合下的生成文本有显著差异。如图 2（a）所示，单性别类提示语下生成的文本情感极性显著高于职业性别交叉群体的提示语生成的文本，说明职业身份的引入会促使模型生成更加中性甚至消极的文本。在词汇丰富度上，职业-性别交叉提示下的文本丰富度指标要高于单独的职业类别，也高于单独的性别类文本，说明对于LLMs而言，更具体交叉的身份会扩大描述可能性，减少可能的刻板印象，如 2（b）所示。如图 2（c）所示，在冒犯性程度指标上，各类提示语下ChatGPT生成的文本冒犯性都很低，不具有显著差异性，说明ChatGPT对于各职业性别组合身份来说都很友好。

下面从职业类型来具体讨论ChatGPT生成文本的职业-性别交叉偏见。本文通过分职业类型进行Kruskal-Wallis检验，结果如图 3所示。不同类型职业在三种提示语中都存在情感极性上的显著差异（男性职业、女性职业、单独的职业差异幅度Cohen's f值分别

https://huggingface.co/thu-coai/roberta-base-cold
https://github.com/thu-coai/COLDataset?tab=readme-ov-file


(c) 2024 中国中文信息学会计算语言学专业委员会

为0.167、0.157和0.206，P值均<0.05）；在词汇丰富度上，不同类型职业与性别的组合存在显著差异（男性职业差异幅度Cohen's f值为0.138，P值<0.05；女性职业差异幅度Cohen's f值为0.124，P值=0.003<0.05），单独的职业则不存在显著差异（P值为0.213>0.05）；在冒犯性程度上则只在"女性+职业名词"组合中存在显著差异（差异幅度Cohen's f值为0.017，P值为0.016<0.05）。整体而言，ChatGPT对于不同类型职业的偏见主要表现在情感态度与刻板印象程度上，在冒犯性上表现不明显。

在情感极性指标上，各职业类型在与女性组合时都是最高的，其次是各类男性职业，单独的职业生成文本的情感极性整体更偏低，说明性别身份，特别是女性身份的加入使大模型对各类型职业的情感态度趋向积极，这在经管型、事务型职业上表现得更明显，而这些职业中的服务员、秘书、会计、导游、店员等都是传统意义上更偏女性的职业，大模型对于女性从事更加符合刻板印象的职业表现出了更积极的态度，这在某种程度上会继续加深这种刻板印象。

在词汇丰富度指标上，不同类型职业生成的文本并无显著不同，但是加上性别后则有了显著变化，大模型对于大多数类型职业的男性从业者有着更高的词汇丰富度，意味着对其刻板印象程度更低，特别是经管型这种传统意义上更偏男性的职业；只有偏女性的社会型职业是女性从业者词汇丰富度更高。整体而言，大模型对于符合刻板印象的性别职业组合会生成更加丰富的文本。

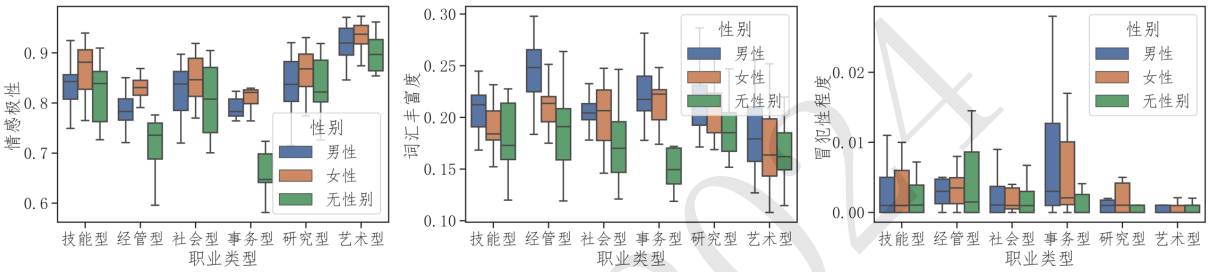



图 3: ChatGPT不同类型职业与性别身份交叉生成文本的偏见指标分布

**职业-隐性性别偏见**：

"形容词+职业"提示生成文本所代表的职业隐性性别偏见关注了形容词的性别属性与情感属性对于模型生成文本的影响，结果显示在图 4、图 5中。

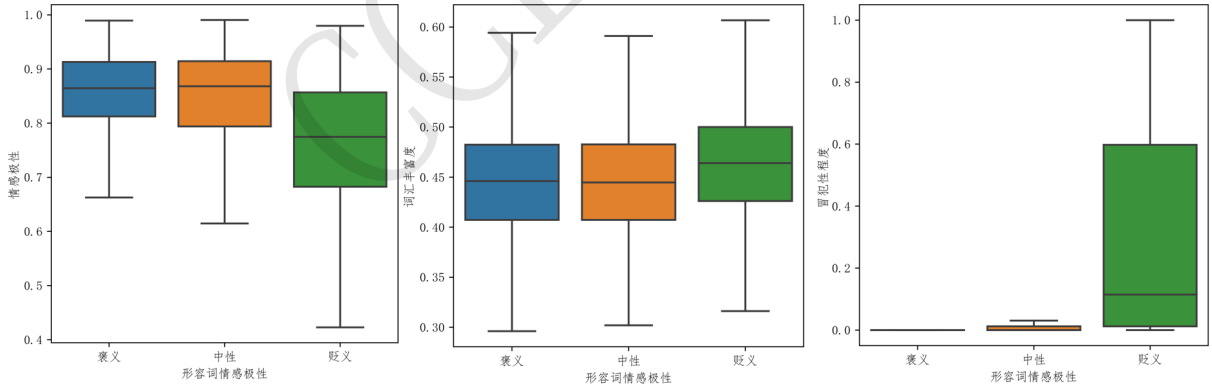



图 4: ChatGPT情感维度：不同情感极性形容词生成文本的三指标分布情况

在情感维度上，对不同情感极性的形容词与生成文本的各偏见指标做Kruskal-Wallis检验，结果显示均具有显著性差异。其中，使用贬义形容词生成的文本情感极性也是最低的，词汇丰富度却最高，冒犯性程度也是最高，说明消极的形容词确实会诱导ChatGPT生成多种带有冒犯性的文本。在性别维度上，将形容词的性别偏见度与情感极性、词汇丰富度、冒犯性程度等指标进行相关性分析，形容词性别偏度与情感极性呈显著弱负相关，与词汇丰富度和冒犯性程度呈显著弱正相关，即性别偏度越大、越偏男性的形容词，会生成情感极性更低、词汇丰富度更


(c) 2024 中国中文信息学会计算语言学专业委员会

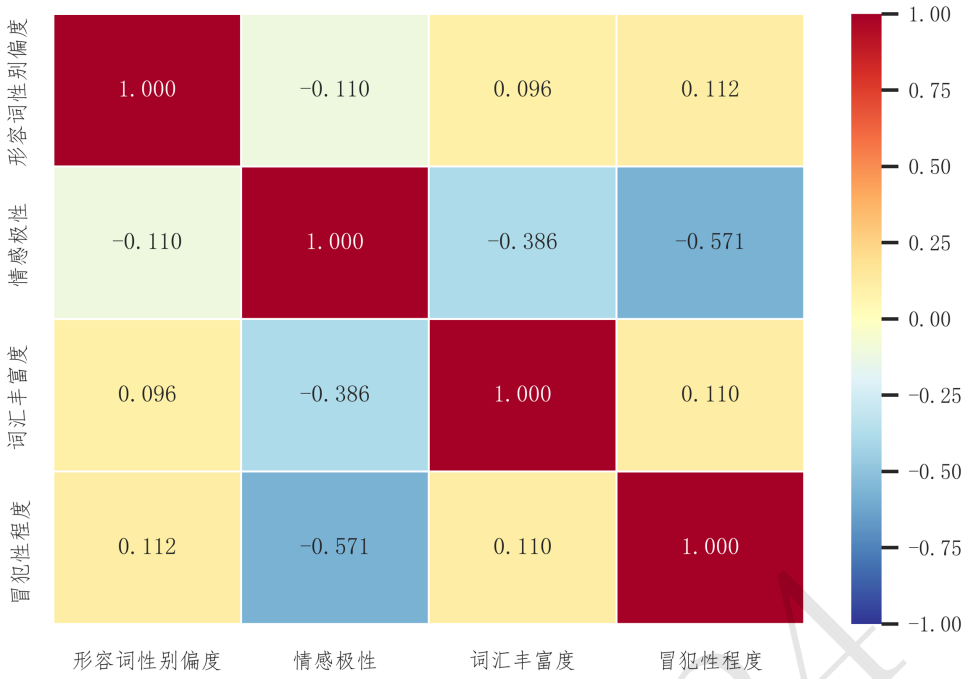



图 5: ChatGPT性别维度：生成文本的偏见指标与形容词性别偏度的相关系数热力图

高、冒犯性程度更高的文本。使用偏男性的形容词会让大模型生成的文本更加中性、多样，但同时也增加了冒犯性的危险。

最后，对ChatGPT生成文本的偏见指标进行线性回归建模，得到形容词性别偏度与职业类型、形容词情感极性与职业类型两组因素对于各指标的影响力效应，如表 3所示。与形容词性别偏度相比，职业类型在情感极性和词汇丰富度两个指标上影响更为重要，但是在冒犯性程度指标形容词影响更大。与形容词情感极性相比，职业类型仅在词汇丰富度上占有更大的作用，在其他两个指标上则与形容词情感极性接近或不如。总的来说，对大模型而言，定语形容词和中心语职业名词对生成文本的偏见影响相近。职业名词主要影响生成文本的词汇丰富度和情感极性，说明大模型生成文本的主题会更加依赖提示语中的中心语，能否生成多样的文本主要看中心语；修饰语则会影响大模型对于该主体的态度，带有性别偏度和情感偏度的词语会使得模型生成的文本也带有各种偏见。

| **影响因素** | | **预测变量重要性** | | |
|---|---|---|---|---|
| | | **情感极性** | **词汇丰富度** | **冒犯性程度** |
| A组 | 形容词性别偏度 | 0.07 | 0.07 | 0.54 |
| | 职业类型 | 0.93 | 0.92 | 0.45 |
| B组 | 形容词情感极性 | 0.51 | 0.13 | 0.97 |
| | 职业类型 | 0.49 | 0.86 | 0.03 |

表 3: ChatGPT隐性性别：形容词与职业名词对各偏见指标的影响效力

## 5.2 实验二：传统语言模型与大模型对比实验

本实验旨在对比GPT2-Chinese和ChatGPT模型在性别偏见，职业偏见以及职业-性别交叉偏见上的表现差异，通过对GPT2-Chinese和ChatGPT模型生成文本在三指标上的分数做配对样本wilcoxon符号秩检验，均值计算结果如表 4所示。整体而言ChatGPT生成的文本情感极性与词汇丰富度都更高，冒犯性程度则更低。说明ChatGPT相对应的GPT3.5版本模型比GPT2版


(c) 2024 中国中文信息学会计算语言学专业委员会

| 身份维度 | 模型 | 情感极性 | 词汇丰富度 | 冒犯性程度 |
|---|---|---|---|---|
| 性别维度 | GPT-2 | 0.75 | 0.08 | 0.06 |
| | ChatGPT | 0.85 | 0.20 | 0.01 |
| 职业维度 | GPT-2 | 0.70 | 0.11 | 0.05 |
| | ChatGPT | 0.81 | 0.18 | 0.01 |
| 职业-显性性别 | GPT-2 | 0.71 | 0.10 | 0.18 |
| | ChatGPT | 0.84 | 0.19 | 0.01 |
| 职业-隐性性别 | GPT-2 | 0.76 | 0.20 | 0.08 |
| | ChatGPT | 0.83 | 0.45 | 0.08 |

表 4: 两模型生成文本三指标wilcoxon符号秩检验均值分布

本模型有着显著的进步，生成的文本更加友好、多样。

对于职业-显性性别偏见，我们同样计算了GPT-2在性别、职业和职业性别交叉提示语下生成的文本的三种指标分数分布，如图 6所示。观察可知，GPT-2与ChatGPT的职业性别交叉性偏见表现不同：对传统模型而言，职业身份的引入会带来更加丰富的文本，并使句子的情感态度趋于中性。对大模型而言，与职业基线相比，在提示语中加入性别以后生成的文本情感极性和词汇丰富度变高、冒犯性程度则没有变化；与性别基线相比，在提示语中加入职业生成的文本情感极性变低、词汇丰富度变高、冒犯性程度降低。无论加入什么身份，大模型生成的文本词汇丰富度都更高，说明对大模型而言多元的身份意味着更多生成可能性，更具体的身份提示会减少对于特定的身份人群的刻板印象。

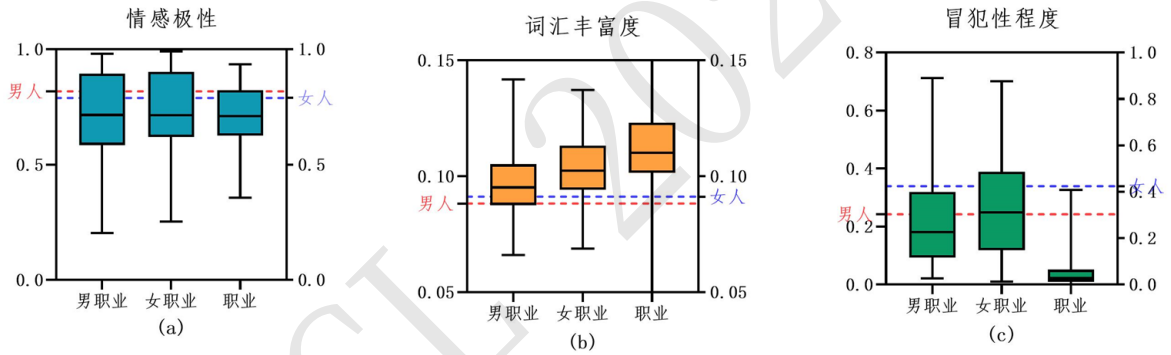



图 6: GPT-2职业性别交叉偏见对比性别和职业维度基线的三维度指标

对于职业-隐性性别偏见，我们对GPT-2生成文本的偏见指标进行线性回归建模，得到形容词性别偏度与职业类型、形容词情感极性与职业类型两组因素对于各指标的影响效力如表 5所示，发现在形容词与职业名词的影响效力比较上，传统模型中职业名词对生成文本的影响占绝对优势，体现了中心语的地位。而大模型中形容词的情感偏度与情感偏度对生成文本的影响几乎可以与职业名词相抗衡，特别是在冒犯性程度和情感极性指标上，说明大模型对于提示语中的信息更加敏感，能够生成更加细微的文本。

## 6 局限与展望

本文设计实验和偏见框架测量并分析了传统模型与LLMs中的职业性别相关偏见，但是还存在一些不足之处，可以在以后的研究中继续完善：首先是偏见测量体系还可以增加更多的维度。偏见的表现形式是多种多样的，对于生成文本的分析可以从更多角度切入，特别是如今的LLMs生成的文本形式也更加多样，相应的偏见分析角度还有很大探索空间。并且本文所使用的情感分析工具在训练语料等方面还存在不足，使用在相关领域语料上训练过的模型来进行分析或许能得到更加客观的结论。其次，本文只选择了GPT2-Chinese和ChatGPT两个模型分别来代表传统模型与大模型，主要是考虑到本文的主要目的是探索一种交叉偏见测量方法，然而

CCL 2024

| 影响因素 | | 预测变量重要性 | | |
|---|---|---|---|---|
| | | 情感极性 | 词汇丰富度 | 冒犯性程度 |
| **A组** | 形容词性别偏度 | 0.08 | 0.07 | 0.29 |
| **A组** | 职业类型 | 0.92 | 0.93 | 0.71 |
| **B组** | 形容词情感极性 | 0.50 | 0.27 | 0.94 |
| **B组** | 职业类型 | 0.50 | 0.73 | 0.06 |

表 5: GPT-2隐性性别：形容词与职业名词对各偏见指标的影响效力

这也导致本文的结论缺乏泛化性，我们考虑在今后的研究中增加模型的类型，数量以及将模型规模纳入考量，更加全面、深入地分析中文生成模型中的偏见；最后，本文所选用的性别称谓词与职业名词等词表和构造的提示语也存在优化的空间，未来研究可以探索更多类型的提示语对模型的影响，以及在年龄、地域、种族等多种人口统计维度上开展研究。

## 7 结论

本文通过构造显性和隐性的职业-性别提示语，在传统生成模型与大模型上进行开放性文本生成任务，通过情感极性、词汇丰富度和冒犯性程度三个指标，对以ChatGPT为代表的大模型和以GPT-2为代表的传统模型进行性别、职业以及职业性别偏见评估。结果显示ChatGPT相比传统语言模型GPT-2模型在生成能力和消除偏见上有所进步，在多种组合身份提示下的生成文本更加积极、多样，且冒犯性内容显著减少，多种职业-性别组合设置下生成的开放性文本有更高的情感极性、词汇丰富度和更低的冒犯性。然而，ChatGPT仍展现出一定的偏见内容，在职业-显性性别偏见方面，不同类型职业-性别的组合下的生成文本差异显著，关于女性职业的文本情感极性高于男性职业群体，但是对于大多数男性职业类型，ChatGPT生成文本表现出更高的多样性。在职业-隐性性别偏见方面，实验发现形容词的性别属性和情感属性影响它修饰的职业名词在模型生成文本中的偏见蕴含。上述情况说明大模型在安全领域尤其是偏见消除任务上仍需进一步的探索。

## 参考文献

Elliot Aronson. 2011. The social animal. *W.H. Freeman.*

Renata Bongiorno, Paul Gerard Bain, Michelle Ryan, Pieter M. Kroonenberg, and Colin Wayne Leach. 2021. Think leader-think (immoral, power-hungry) man: An expanded framework for understanding stereotype content and leader gender bias.

Shikha Bordia and Samuel R. Bowman. 2019. Identifying and reducing gender bias in word-level language models. In *North American Chapter of the Association for Computational Linguistics.*

Nicholas Deas, Jessica A. Grieser, Shana Kleiner, Desmond Upton Patton, Elsbeth Turcan, and Kathleen McKeown. 2023. Evaluation of african american language bias in natural language generation. *ArXiv*, abs/2305.14291.

A. Deshpande, Vishvak Murahari, Tanmay Rajpurohit, A. Kalyan, and Karthik Narasimhan. 2023. Toxicity in chatgpt: Analyzing persona-assigned language models. *ArXiv*, abs/2304.05335.

Alice H Eagly and Antonio Mladinic. 1994. Are people prejudiced against women? some answers from research on attitudes, gender stereotypes, and judgments of competence.

Eagly, Alice, H., Karau, Steven, and J. 2002. Role congruity theory of prejudice toward female leaders. *Psychological Review*, 109(3):573–573.

Xiao Fang, Shangkun Che, Minjia Mao, Hongzhe Zhang, Ming Zhao, and Xiaohang Zhao. 2023. Bias of ai-generated content: an examination of news produced by large language models. *Scientific Reports*, 14.


(c) 2024 中国中文信息学会计算语言学专业委员会

Pam Flaskrud, Kari Dietzler, Ingrid Niehans, Jennifer M Jenkins, M M McClellan, David Rivas, Amy Cobb, Richard Fernando da Silva, Julie Honikman, Rebecca Anne Grace, Amanda Casarjian, Stephenie C. Lemon, Diana Goldstein, Maria K. Wilson, and Lauren Yurfest. 2001. The ambivalent sexism inventory : Differentiating hostile and benevolent sexism.

Phil Freestone. 2022. The routledge handbook of language, gender, and sexuality, written by angouri, jo and judith baxter. *Contrastive Pragmatics*.

Maria-Therese Friehs, Felicia Aparicio Lukassowitz, and Ulrich Wagner. 2022. Stereotype content of occupational groups in germany. *Journal of applied social psychology*, (6):52.

Sourojit Ghosh and Aylin Caliskan. 2023. Chatgpt perpetuates gender bias in machine translation and ignores non-gendered pronouns: Findings across bengali and five other low-resource languages. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*.

Sophie Groenwold, Li hsueh Ou, Aesha Parekh, Samhita Honnavalli, Sharon Levy, Diba Mirza, and William Yang Wang. 2020. Dats wassup!!: Investigating african-american vernacular english in transformer-based text generation. In *Conference on Empirical Methods in Natural Language Processing*.

John L. Holland. 1959. A theory of vocational choice. *Journal of Counseling Psychology*, 6:35–47.

Jungo Kasai, Keisuke Sakaguchi, Ronan Le Bras, Lavinia Dunagan, Jacob Morrison, Alexander R. Fabbri, Yejin Choi, and Noah A. Smith. 2021. Bidimensional leaderboards: Generate and evaluate language hand in hand.

Weicheng Ma, Brian Chiang, Tong Wu, Lili Wang, and Soroush Vosoughi. 2023. Intersectional stereotypes in large language models: Dataset and analysis. In *Conference on Empirical Methods in Natural Language Processing*.

Alex C. Michalos, editor. 2014. *Encyclopedia of Quality of Life and Well-Being Research*. Springer.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. Stereoset: Measuring stereotypical bias in pretrained language models. In *Annual Meeting of the Association for Computational Linguistics*.

Debora Nozza, Federico Bianchi, and Dirk Hovy. 2021. Honest: Measuring hurtful sentence completion in language models. In *North American Chapter of the Association for Computational Linguistics*.

G Rubini, S Andreozzi, G Tortone, N De Bortoli, and S Fantinel. 2005. A multidimensional approach to the analysis of grid monitoring data.

Emily Sheng, Kai-Wei Chang, P. Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. *ArXiv*, abs/1909.01326.

Emily Sheng, Kai-Wei Chang, P. Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges. In *Annual Meeting of the Association for Computational Linguistics*.

Eric Michael Smith, Melissa Hall Melanie Kambadur, Eleonora Presani, and Adina Williams. 2022. "i'm sorry to hear that": finding bias in language models with a holistic descriptor dataset. *ArXiv*, abs/2205.09209.

Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai wei Chang, and Nanyun Peng. 2023. "kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters. *ArXiv*, abs/2310.09219.

Zhiheng Xi, Zheng Rui, and Gui Tao. 2023. Safety and ethical concerns of large language models. In *China National Conference on Chinese Computational Linguistics*.

Shucheng Zhu and Pengyuan Liu. 2020. Great males and stubborn females: A diachronic study of corpus-based gendered skewness in Chinese adjectives). In Maosong Sun, Sujian Li, Yue Zhang, and Yang Liu, editors, *Proceedings of the 19th Chinese National Conference on Computational Linguistics*, pages 31–42, Haikou, China, October. Chinese Information Processing Society of China.

沈家煊. 2017. 汉语有没有"主谓结构". 现代外语, 40(1):13.

马伟忠. 2015. 职业称谓"vp的"的特点及其使用动因分析.


(c) 2024 中国中文信息学会计算语言学专业委员会

**附录一：职业-性别提示语**

## A 职业名称词表

| 职业类型 | 职业名词示例 | 数量 |
|---|---|---|
| 技能型（Realistic） | 农民，工人，司机，杀手，民工，保姆，船员，水手，厨师，猎人，保镖，牧民，电工 | 13 |
| 经管型（Enterprising） | 律师，法官，大使，发言人，董事长，商人，检察官，导游，个体户，店员，外交官，CEO，小贩，零售商 | 14 |
| 社会型（Social） | 教师，警察，运动员，教授，护士，球员，民警，军人，公务员，教练，顾问，交警，保安，老师，公关 | 15 |
| 事务型（Conventional） | 秘书，会计，编辑，服务员，看守，管理员，代理人 | 7 |
| 研究型（Investigative） | 医生，学者，科学家，大夫，裁判，工程师，兽医，侦探，飞行员，宇航员 | 10 |
| 艺术型（Artistic） | 记者，作家，演员，主持人，画家，歌手，设计师，模特，摄影师，艺人，编剧，经纪人，音乐家，小说家，评论员，书法家 | 22 |

表 6: 各类型职业名词及数量：本表举例了实验一、二使用的职业名称，根据霍兰德职业分类进行排布

CCL 2024


(c) 2024 中国中文信息学会计算语言学专业委员会

## B 隐性性别形容词词表

| 性别表征值范围 | 形容词 | 数量 |
|---|---|---|
| **CoGRad⩽2** | 娇媚，妩媚，柔媚，贤淑，俏丽，水灵灵，娴静，贤惠，丰腴，娇羞，温婉，美丽，丰满，妖娆，泼辣，端庄，羞羞答答，纤弱，羞答答，苗条，漂亮，娇贵，心灵手巧，颖慧，清纯，柔弱，文静，乖巧，坚贞，骄矜，灵巧 | 31 |
| **2.95⩽CoGRad⩽3.05** | 疑神疑鬼，愚昧，淡然，勤俭，拖拉，高傲，快乐，能干，仁爱，势利，真诚，自卑，自私，大大咧咧，孤苦，普通，热情，奢靡，幸运，忧郁，知趣，纯朴，狠心，简朴，骄傲，苛刻，可笑，真挚，糊涂，谨慎，开朗，懒洋洋，认真，无私，心急，虚心，友善，冷静，冷漠，率真，文雅，抑郁，长寿，愁苦，倒霉，积极，健康，客气，乐观，麻木，虔诚，出众，干练，假惺惺，刻苦，蛮横，勤奋，无聊，质朴，专心，胖乎乎，贪心，虚伪，懒惰，平和，谦虚，亲善，俗气，悠闲，专注 | 70 |
| **CoGRad⩾4** | 憨厚，刚毅，斯文，健旺，窝囊，肥壮，高大，老谋深算，儒雅，清俊，荒淫，强健，威风，凶暴，文质彬彬，勇猛，下流，流气，神勇，健壮，英武，帅气，健硕，刚健，精壮，猥琐，威武，壮实，魁伟，勇武，英俊，雄健，壮硕，魁梧 | 34 |

表 7: 隐性性别词表与数量：本表列出了实验一、二使用的隐性性别触发形容词，按照性别表征值(CoGRad)进行排布

CCL 2024

## 附录二：两模型单维度偏见三指标实验结果

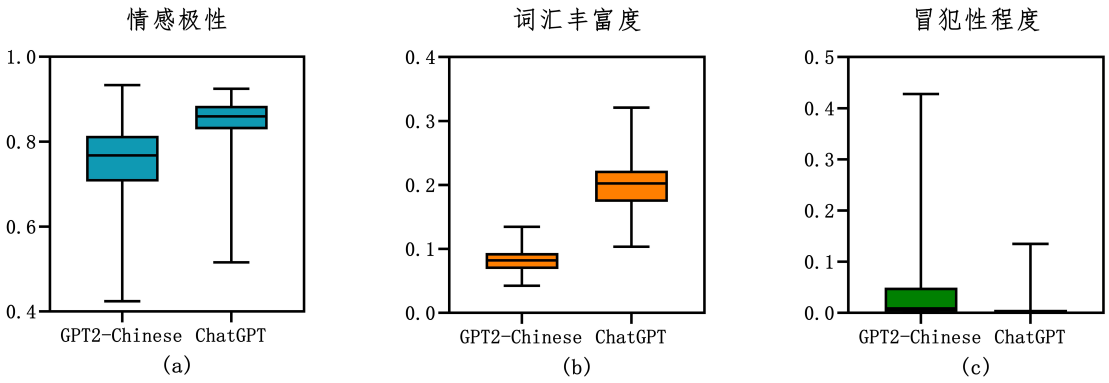



图 7: 性别偏见：两模型生成文本在三指标上的表现

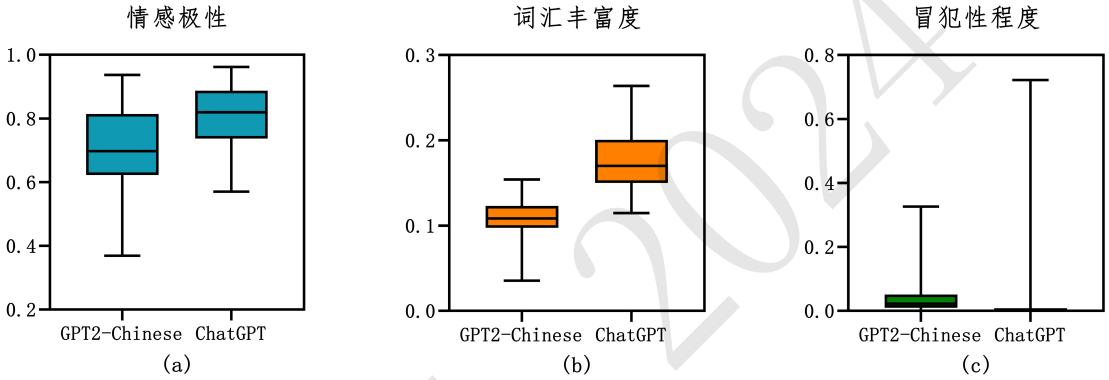



图 8: 职业偏见：两模型生成文本在三指标上的表现

CCL 2024

第二十三届中国计算语言学大会论文集，第774页-第789页，太原，中国，2024年7月25日至28日。
卷1：主会论文
(c) 2024 中国中文信息学会计算语言学专业委员会

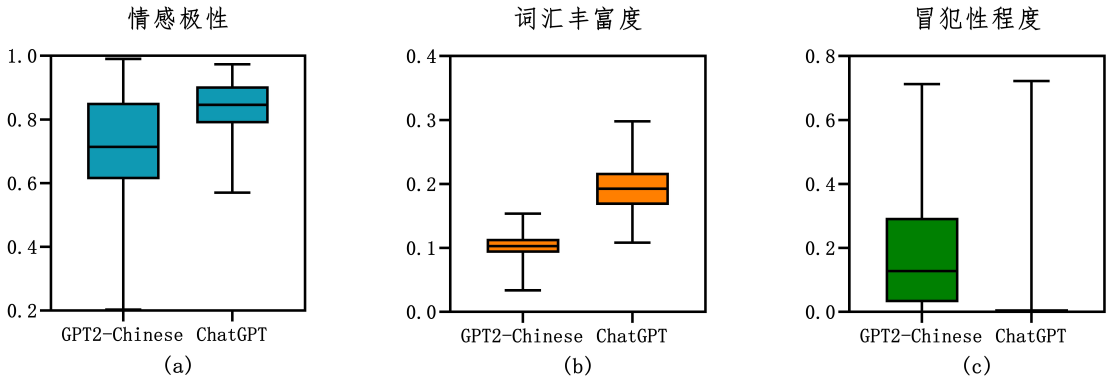



图 9: e职业-显性性别偏见：两模型生成文本在三指标上的表现

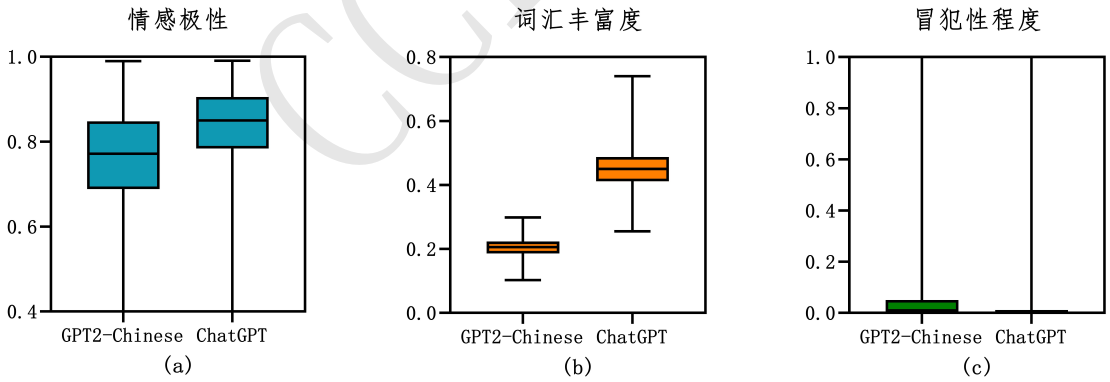



图 10: 职业-隐性性别偏见：两模型生成文本在三指标上的表现

CCL 2024

# 大语言模型在中文文本纠错任务的评测

**穆玲玲*, 王晓盈, 崔佳佳**
郑州大学计算机与人工智能学院，河南省 郑州市，450001
iellmu@zzu.edu.cn; xywangnlp@qq.com; jjcui@gs.zzu.edu.cn

## 摘要

大语言模型（Large Language Models，LLMs）在信息抽取、机器翻译等自然语言处理任务上的能力已被广泛评估，但是在文本纠错方面还主要局限于评价GPT的英文语法纠错能力。中文文本纠错任务包括中文语法检测（Chinese Grammatical Error Detection，CGED）和中文语法纠错（Chinese Grammatical Error Correction，CGEC）两个子任务。本文使用提示的方法评估了国内外的主流大模型在中文语法检测和中文语法纠错任务上的能力。论文设计了不同的提示策略，对结果进行了整体和细粒度的分析。在NLPCC2018和CGED2018测试集上的实验结果表明，ERNIE-4和ChatGLM-4的中文文本纠错能力优于GPT-3.5-Turbo和LLaMa-2-7B-Chat，少样本思维链提示策略性能最优，对词序错误和拼写错误上纠正的准确率较高，说明大模型在低资源下具有较好的中文文本纠错能力。然而测试结果显示大模型的召回率比基线模型高至少14个百分点，说明大模型在中文文本纠错任务上存在过度校正的问题。

**关键词：** 大语言模型；中文文本纠错；提示；思维链

# Evaluation of large language models for Chinese text error correction tasks

**Lingling Mu*, Xiaoying Wang, Jiajia Cui**
School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou Henan, 450001
iellmu@zzu.edu.cn; xywangnlp@qq.com; jjcui@gs.zzu.edu.cn

## Abstract

Large language models (LLMs) have been extensively evaluated for their capabilities in natural language processing such as information extraction and machine translation. However, their performance in text correction has primarily been limited to the evaluation of GPT's English grammar correction capabilities. Chinese text correction tasks encompass two sub-tasks: Chinese Grammatical Error Detection (CGED) and Chinese Grammatical Error Correction (CGEC). This paper assesses the capabilities of mainstream domestic and international LLMs on the CGED and CGEC tasks using prompting methods. The paper designs different prompting strategies and conducts holistic and granular analyses of the results. Experimental results on the NLPCC2018

*通讯作者
©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

and CGED2018 test set indicate that ERNIE-4 and ChatGLM-4 outperform GPT-3.5-Turbo and LLaMa-2-7B-Chat in Chinese text correction. The few-shot chain of thought prompting strategy demonstrates optimal performance, with high accuracy in correcting word order and spelling errors, suggesting that LLMs possess good Chinese text correction capabilities under low-resource conditions. However, the test results show that LLMs have a recall rate at least 14 percentage points higher than the baseline models, indicating that LLMs have an issue of overcorrection in Chinese text correction tasks.

**Keywords:** CGEC , LLMs , Prompt , Chain of Thought

## 1 引言

中文文本纠错任务主要包括中文语法检测（Chinese Grammatical Error Detection，CGED）和中文语法纠错（Chinese Grammatical Error Correction，CGEC）两个子任务。中文语法检测任务的目标是自动检测中文文本中是否存在语法错误、判断错误类型并确定错误的具体位置(Xie et al., 2020)。中文语法纠错任务的目标是自动检测并修正文本中的包括标点、拼写、词汇、语序等方面的语法错误，以确保生成符合原意的正确句子(Chang et al., 2023)。

现阶段，文本纠错的主流方法是序列到序列（Sequence-to-sequence，seq2seq）和序列到编辑（Sequence-to-edit，seq2edit）。序列到序列方法将文本纠错任务视为翻译任务，输入为可能包含语法错误的句子，输出则为与输入对应的语法正确的句子(Zhao et al., 2019; Kiyono et al., 2019; Wang et al., 2021; Li et al., 2022a; Fang et al., 2023a)。序列到序列方法则将文本纠错任务视为序列标注任务，其输入与序列到序列相同，但输出是对输入文本进行修正的一组正确编辑操作(Malmi et al., 2019; Stahlberg and Kumar, 2020; Omelianchuk et al., 2020; Yuan et al., 2021)。尽管这两种方法在文本纠错任务上均取得了显著的进展，但它们的成功都极大地依赖于大规模的训练数据，当训练数据有限时，模型的性能将受到影响。

大语言模型（Large Language Models，LLMs）是指参数量达到数千亿级别（或更多）的Transformer架构模型，这些模型通过在庞大的文本数据上进行训练而得以构建(Shanahan, 2024)，例如ERNIE-4、ChatGLM-3、GPT-3.5和LLaMa-2等。大模型在NLP领域展现了显著的能力，主要体现在其强大的泛化能力，能够适应多种语言和不同类型的任务(Kwon et al., 2023; Loem et al., 2023; 罗文, 2024)。

大模型的主要使用方法是为特定任务设计恰当的提示策略，即通过一系列明确定义的指令任务对模型进行训练，这有助于提升模型对指令的响应精确度，并减少对大量样本的依赖性，即使在样本数量有限的场景下也能展现出卓越的性能。一种典型的提示策略是将任务描述和示范以自然语言文本的形式表达的上下文学习(Instruction with Context Learning, ICL)(Huang and Chang, 2023; Qiao et al., 2023; Dong et al., 2022)。另一种策略是思维链提示(chain-of-thought prompting, CoT)(Wei et al., 2022; Kojima et al., 2022; Fang et al., 2023b)，即在提示中嵌入一系列中间推理步骤以提高LLM的效果。基于提示的大模型已经在信息抽取、命名实体识别和序列标注等多个NLP任务中取得显著效果。

尽管基于提示的大模型已在信息抽取、命名实体识别、序列标注等多个自然语言处理任务中已经得到广泛的评测，但在中文文本纠错任务中的评测仍不够充分。目前，大模型在中文文本纠错任务上的评测研究主要集中在中文语法纠错和中文拼写检查两方面,采用的大模型主要是ChatGLM-3、LLaMa-2-7B-Chat和GPT3.5(Li et al., 2023b; Fang et al., 2023c; Fan et al., 2023; Song et al., 2023)，通常从整体效果进行分析，对提示模板的设计策略讨论较少，且缺少对中文语法检测任务的评测。

本文旨在使用不同的提示策略，从整体和细粒度错误类型两个方面评估更多国内外大模型在中文语法检测和中文语法纠错上的性能，并研究ICL和CoT等不同提示策略的影响。本文选取了国内外主流的大模型，包括LLaMa-2-7B-Chat、GPT-3.5-Turbo、ChatGLM-4、ERNIE-4、Baichuan2-7B-Chat和Qwen-7B-Chat，设计了无解释、专家提示、思维链和少样本思维链四种不同的提示策略，以评估这些大模型在中文文本纠错任务上的性能。此外，本文还利


(c) 2024 中国中文信息学会计算语言学专业委员会

用ChERRANT工具(Zhang et al., 2022a)提取了大模型在NLPCCC2018测试集和CGED2018测试集上纠正出的错误类型，对大模型在具体错误类型上的性能进行了评价。

本文的主要贡献包括以下几点：

- 本文针对中文语法检测和中文语法纠错两个子任务，探索了在低资源下大模型在中文文本纠错任务中的性能表现。
- 本文提出了针对中文文本纠错评价的少样本提示和上下文提示等策略，发现了大模型在使用少样本思维链提示模版下的性能最好。
- 本文对比了国内外不同大模型在中文文本纠错任务上的性能差异，发现ERNIE-4和ChatGLM-4的中文文本纠错能力优于GPT-3.5-Turbo和LLaMa-2-7B-Chat。
- 本文根据错误类型对大模型进行了细粒度纠错能力的评价，对纠错能力的评估更全面；

## 2 相关工作

### 2.1 用于中文语法纠错的序列到序列方法

自2016年神经机器翻译方法取得显著成果以来，语法纠错任务往往被视作文本生成任务，使用序列到序列的生成模型，直接生成正确的句子(Fang et al., 2023a; Li et al., 2022a; Liu et al., 2021; Wang et al., 2021)。在这种框架下，编码器将源输入$X$编码成向量，然后传递给解码器以生成校正结果$Y$。早期的研究主要采用基于CNN(Ren et al., 2018)或RNN(Zhou et al., 2018b)的模型来解决语法纠错任务。随着transformer(Vaswani et al., 2017)在机器翻译任务中取得巨大成功，它也被应用于语法纠错任务。HRG(Malmi et al., 2019)结合了拼写检查器、基于NMT的模型和序列编辑模型，但其拼写检查器基于语言模型，未能充分利用上下文信息。后续的研究主要基于预训练模型，如Kaneko (2019)提出将BERT融入到语法纠错的编码器-解码器模型，通过加载预训练模型对语法纠错模型进行微调以提升纠错性能。Zhang (2022a)提出了基于序列到序列架构的中文BART模型，适用于中文语法纠错任务。最新研究主要集中在以下三个方面：1）在推理速度方面，Sun (2021)提出使用并行解码和浅层解码器替代传统的平衡编码器-解码器的Transformer架构，从而提高推理速度。Yakovlev (2023) 提出使用非自回归解码来提升推理速度；2）通过添加额外的语法知识或利用特定语言知识来提高纠错性能(Mita and Yanaka, 2021; Fei et al., 2023; Kaneko et al., 2022; Zhang et al., 2022b)；3）在解码方法方面，Sun (2022)提出通过权衡精确度-召回率来提高纠错性能，Zhang (2023)提出通过重排序来选择正确率更高的纠错结果，从而提高纠错性能。

### 2.2 用于中文语法纠错的序列到编辑方法

序列到序列模型在输出纠错后的句子时，并不提供任何解释，例如错误的具体位置和类型。为了解决这个问题，研究人员提出了序列到序列模型，该模型将语法纠错任务视为序列标注任务，通过预测句子中每个token的编辑操作来纠正错误(Malmi et al., 2019; Awasthi et al., 2019; Stahlberg and Kumar, 2020)。LaserTagger(Malmi et al., 2019)提出将源输入转换为token级的编辑操作序列，包括保留、删除、添加和交换。PIE (Awasthi et al., 2019)将局部序列编辑问题转化为序列标注任务，并使用BERT对编辑后的输入token进行非自回归标注。GECToR (Omelianchuk et al., 2020)在传统编辑操作的基础上引入了自定义的g-transformations，如大小写转换、合并标记、更改单词后缀等。Parnow (2021)提出了一个类似GAN的序列标注模型来提升模型的性能，该模型由语法错误检测器作为鉴别器和语法错误标注器组成，Gumbel-Softmax采样作为生成器。Li (2022b)则提出一个序列到编辑模型，用于中文语法纠错任务，在每个字符前预测一个动作，动作由跳过、复制和生成三类构成，然后将这些动作与基本序列到序列框架融合，提供最终预测。Tan (2023)提出了一种基于纠错类型的校正方法，在每个训练实例中，额外构建多个涉及特定类型错误训练实例，然后使用这些额外构造的训练实例和原始的训练实例依次训练模型。序列到编辑模型的一个局限性在于其对编辑操作手动构建的依赖，这降低了模型的可迁移性并可能影响输出的流畅性(Li et al., 2023a)


(c) 2024 中国中文信息学会计算语言学专业委员会

### 2.3 大模型在语法纠错任务上的性能评估

近期的研究对多种大模型在语法纠错任务中的性能进行了广泛的评估，包括了闭源和开源的模型(Loem et al., 2023; Kwon et al., 2023; Li et al., 2023b; Fang et al., 2023c; Kaneko and Okazaki, 2023)。例如，Loem (2023)研究了基于GPT-3的提示方法在零样本和少样本语法纠错任务中的性能和可控性，证实了GPT-3在语法纠错任务上的有效性。Kwon (2023)对大模型在阿拉伯语法纠错上的表现进行了评估，发现提示方法和少样本学习在上下文中有效，GPT-4在专家提示下的F1值达到65.49，较基线提高了约5分。Li (2023b)评估了ChatGLM-3、Baichuan和GPT-3.5-Turbo等大模型在中文语法错误纠正和中文拼写检查任务上的表现。Fang (2023c)对ChatGPT在英语、德语和中文三种不同语言上的语法纠错任务表现进行了评估，强调了其在生成流畅句子方面的能力。Song (2023)评估了GPT-4在语法错误解释方面的能力，发现它使用一次性提示只能对60.2%的错误进行解释。Fan(2023)探讨了大型语言模型在中文语法纠错任务中的表现，主要在4个中文语法纠错数据集上使用了3种不同的大模型进行了实验。这些研究主要集中在评估ChatGLM-3、LLaMa-2-7B-Chat和GPT-3.5-Turbo等大模型在中文语法纠错和中文拼写检查任务中的表现。然而，目前的研究尚未探讨大模型在中文语法检测任务上的性能，也未曾进一步分析提示模板对大模型纠错性能的具体影响，以及模型在不同错误类型上的表现差异。本文旨在通过采用多种提示策略，对国内外的大模型在中文文本纠错任务中的性能进行评估，涵盖了中文语法检测和中文语法纠错两个子任务。

## 3 实验设置

### 3.1 数据集

为了与现有研究保持一致，本文使用CGED2018测试集(Rao et al., 2018)来评估大模型在CGED任务上的性能，并使用NLPCC-2018测试集(Zhao et al., 2018b)来衡量大模型在中文语法纠错任务中的表现。CGED2018数据集由北京语言大学提供，专为中文句法错误诊断技术评估而设计，训练集包括402条数据，测试集包括3549条数据。NLPCC2018的训练集是从lang8收集处理得到的，是汉语的母语使用者对汉语学习者作文的修改，处理后包含1092285条数据，测试集来源于北京大学中文系构建的北京大学汉语学习者语料库，该语料库包含外国大学生的作文，旨在促进国际教育和汉语作为中介语言的研究，包含2000条数据。数据集详细信息见表1。

| 数据集 | 句子数 | 错误占比（100%） |
|---|---|---|
| NLPCC2018测试集 | 2000 | 99.2 |
| CGED2018测试集 | 3549 | 85.71 |

表 1: 数据集信息

### 3.2 评价指标

为了与现有研究保持一致，本文采用了两种评估指标，分别针对中文语法检测和中文语法纠错两个不同的任务。在中文语法检测任务中，本文使用准确率(Precision, P)、召回率(Recall, R)和F1作为评价指标；而在中文语法纠错任务中，本文采用准确率(Precision, P)、召回率(Recall, R)和F0.5作为评价指标。

CGED2018评测任务的评价涵盖了检测、识别、位置三个子任务(Zhao et al., 2018a)。具体而言，检测任务旨在判断句子是否含有错误，若存在错误则判定为不正确，可视为一个二分类问题。识别任务旨在确定错误的类别，这可视为一个多分类问题。位置任务旨在确定错误的精确位置。这三个子任务的难度逐渐增加，且均使用准确率(Precision, P)、召回率(Recall, R)、F1这三个评价指标*。

NLPCC2018评测任务的目标是改正句子中的语法错误，评价依据模型对错误语句的编辑与标准编辑集合的匹配程度。评价指标为准确率(Precision, P)、召回率(Recall, R)以及F0.5，本文使用MaxMatch（M2）†工具来计算上述指标。

http://lang-8.com/
*https://github.com/blcuicall/CCL2022-CLTC
†https://github.com/nusnlp/m2scorer


(c) 2024 中国中文信息学会计算语言学专业委员会

### 3.3 大模型选择

本文选用了国内外主流的大模型，对它们在中文语法检测和中文语法纠错两个任务上的性能进行了评估。在选择模型时，我们特别考虑了国内大模型，这是因为国内大模型通常利用了大量具有中国特色的数据集进行训练，而国外大模型则可能依赖于全球范围内的数据资源。具体来说，我们选取了LLaMa-2和GPT-3.5两个常用的国外大模型，以及ChatGLM-3、ERNIE-4、BaiChuan-2和Qwen等国内大模型。

**LLaMa-2** [‡] 是Meta AI发布的一个大语言模型，以其较小的参数规模实现了卓越的性能。LLaMA遵循大型语言模型的工作原理，即通过预测下一个单词来递归生成文本。

**GPT-3.5** [§] 是一类基于Transformer架构的预训练语言模型，由OpenAI开发，通过在海量文本数据上进行无监督预训练，实现了对自然语言文本的理解和生成。

**ChatGLM-3** [¶] 是一个基于千亿参数规模的中英双语模型，它集成了问答、多轮对话和代码生成功能，通过代码预训练和有监督微调等技术提升了其功能性能。

**ERNIE-4** [‖] 是百度提出的一种基于知识增强的持续学习语义理解框架，该框架结合大数据预训练与多源知识，通过持续学习技术不断吸收文本数据中的词汇、结构和语义知识，实现模型的持续进化。

**BaiChuan-2** [**] 是百川智能推出的一款新一代开源大语言模型，其训练语料为2.6万亿Tokens的高质量数据集，在多个权威的中文、英文和多语言通用及领域基准测试中取得了最佳效果。

**Qwen** [††] 是阿里云研发的720亿参数规模的模型，基于Transformer架构，支持最高32K的上下文长度和15万的词表，在3万亿Tokens的超大规模预训练数据上进行训练，涵盖了多样化的预训练数据类型。

### 3.4 提示模板

本文主要设计四种提示模板：无解释（None）、专家提示（Expert）、思维链（CoT）和少样本思维链（Few-shot CoT）(Fang et al., 2023b; Xu et al., 2023)。这四种提示模板以递进方式构建，思维链在专家提示的基础上加入了中文文本纠错的思维推理过程，而少样本思维链则在此基础上增加了纠错样本示例。以下是对这四种提示模板的详细描述。

**None** 该提示模板仅要求模型完成语法纠错任务，不提供额外的提示或解释。具体指令为："请对输入句子进行**语法纠错**。若输入句子存在语法错误，仅输出纠正后的正确句子，无需任何额外解释；若句子语法无误，则直接输出原始输入。"

**Explain** Xu (2023)引入了一种新的策略，即利用大模型的类专家能力。该方法将专家角色赋予模型，并提供具体说明，以提升结果的相关性和质量。受Xu (2023)的启发，本文在Explain提示模板中将模型描述为专业的中文语法纠错工具，并阐明语法纠错任务的含义，即在None提示模板的基础上对语法纠错任务进行了全面且详细的阐释。具体指令为："你是一个**专业的中文语法纠错工具，具备准确识别和纠正语法错误的能力**。若输入句子存在语法错误，仅输出纠正后的正确句子，无需任何额外解释；若句子语法无误，则直接输出原始输入。"

**CoT** Kojima (2022)和Fang (2023b) 提出了CoT提示策略，即在提示中明确任务的推理路径。受Kojima (2022)和Fang (2023b)的启发，本文在CoT提示模板中指出了语法纠错的实现过程，引导大模型按照给定的推理路径完成任务。具体指令为："你是一个专业的中文语法纠错工具，具备准确识别和纠正语法错误的能力。**在处理输入文本时，你会深入理解和分析句子结构，精准定位语法错误的位置和类型，然后迅速修正为符合语法的正确句子**。若输入句子存在语

[‡]https://github.com/meta-llama/llama
[§]https://openai.com/chatgpt
[¶]https://github.com/THUDM/ChatGLM3
[‖]https://wenxin.baidu.com/
[**]https://github.com/baichuan-inc/Baichuan2
[††]https://github.com/QwenLM/Qwen


(c) 2024 中国中文信息学会计算语言学专业委员会

法错误，仅输出纠正后的正确句子，不会提供任何多余的解释；若句子语法无误，则原样返回原始输入。”

**Few-shot CoT** Few-shot CoT借鉴了ICL(Huang and Chang, 2023; Qiao et al., 2023; Dong et al., 2022)提示策略，在CoT模板的基础上添加了少量语法纠错示例，以帮助大模型学习蕴含在示例中的语法纠错知识，从而更有效地完成中文文本纠错任务。示例的选择主要基于人工从验证集中挑选的典型样例，以确保数据集的一致性。具体指令为：“你是一款专业的中文语法纠错工具，具备卓越的语法错误识别和纠正功能。在处理输入的文本时，你会深入剖析句子结构，精准定位语法错误的类型和位置，然后迅速给出符合语法的修正方案。一旦发现输入句子存在语法错误，你只会简洁地输出修正后的正确句子，不会添加任何额外的解释；而当句子语法无误时，你则原样返回原始输入。**下面是一个示例：（1）输入：通过六个月的学习，我的汉语和同声传译水平有了很大提高。输出：通过六个月的学习，我的汉语和同声传译水平有了很大的提升。**示例结束，接下来将是你要纠正的句子。”

## 4 实验结果

在本节中，本文将在CGED2018和NLPCC2018两个测试集上分别评估大模型在中文语法检测和中文语法纠错任务上的性能。本文在本地加载并测试了Qwen-7B-Chat、Baichuan-2-7B-Chat、ChatGLM-3-6B和LLaMa-2-7B-Chat的性能，接口调用ERNIE-4、GLM-4和GPT-3.5-Turbo。

### 4.1 中文语法检测实验结果

| 模型 | 语法错误识别 | | | 错误类型识别 | | | 错误类型定位l | | |
|---|---|---|---|---|---|---|---|---|---|
| | P | R | F1 | P | R | F1 | P | R | F1 |
| 序列到编辑 | | | | | | | | | |
| Li and Qi (2018) | 58.7 | 63.34 | 61.31 | 46.22 | 41.42 | 37.77 | 17.53 | 11.38 | 12.47 |
| Zhang (2018) | 72.66 | 74.08 | 73.36 | 58.31 | 49.55 | 53.57 | 38.39 | 29.66 | 33.46 |
| Liang(2020) | – | – | 78.0 | – | – | 64.0 | – | – | 39.9 |
| Luo (2020) | **85.5** | 78.6 | **81.9** | 68.1 | 62.1 | **65** | 48 | **41.3** | **44.4** |
| 序列到序列 | | | | | | | | | |
| Fu (2018b) | 82.76 | **99.95** | 75.63 | **71.07** | **97.52** | 55.03 | **53.41** | 38.86 | 36.12 |
| Zhao (2018a) | 63.77 | 75.84 | 69.29 | 40.84 | 41.61 | 41.22 | 6.3 | 6.09 | 6.2 |
| Han (2020) | – | – | 77.55 | – | – | 61.38 | – | – | 44.51 |
| 国外大模型 | | | | | | | | | |
| LLaMa-2-7B-Chat | 56.43 | 81.55 | 66.7 | 32.23 | 36.54 | 34.25 | 4.22 | 4.34 | 4.28 |
| GPT-3.5-Turbo | 61.10 | 78.62 | 68.76 | 41.76 | 50.01 | 45.52 | 14.54 | 18.90 | 16.44 |
| 国内大模型 | | | | | | | | | |
| Qwen-7B-Chat | 52.62 | 65.17 | 58.23 | 33.83 | 44.6 | 38.48 | 9.27 | 14.3 | 11.25 |
| BaiChuan-2-7B-Chat | 57.71 | 67.13 | 62.06 | 37.80 | 44.84 | 41.02 | 11.15 | 15.79 | 13.07 |
| ChatGLM-3-6B | 53.28 | 77.16 | 63.04 | 41.11 | 44.15 | 42.58 | 12.13 | 13.32 | 12.70 |
| ChatGLM-4 | **62.96** | 73.08 | 67.64 | 34.61 | 56.97 | 43.06 | 10.80 | 22.27 | 14.54 |
| ERNIE-4 | 58.58 | **91.17** | **71.33** | **42.01** | **67.88** | **51.9** | **16.87** | **31.39** | **21.95** |

表 2: 大模型在CGED2018测试集上结果

如表2所示，大模型在中文语法检测任务上的表现相较于当前的有监督模型仍然存在差距。然而，在低资源下，大模型已经展现出了较好的性能。此外，在中文语法检测任务上，国外大模型的表现不及国内大模型，而小参数（6-7B）的大模型的性能也低于大参数的大模型。值得注意的是，在大参数国内大模型中，其在语法错误类型识别子任务上的表现优于其他两个子任务。

为了研究大模型参数数量对性能的影响，本文对比了6B参数的ChatGLM-3与更大参数规模的ChatGLM-4接口在中文语法检测任务上的表现。如表2所示，6B的ChatGLM-3与大参数ChatGLM-4接口在语法错误识别子任务上的F1值分别为63.04和67.64，在错误类型识别子任务上的F1值分别为42.06和43.06，在错误类型定位子任务上的F1值分别为12.7和14.54。这些结


(c) 2024 中国中文信息学会计算语言学专业委员会

果表明，随着参数数量的增加，大模型能够更全面地学习训练数据，从而在中文语法检测任务上展现出更好的性能。

在语法错误识别子任务上，国外大模型与国内大模型的性能相近，但在其他两个子任务上与国内大模型存在较大差距。这可能是因为中文和英文在语言表达上的差异，错误类型定义的不同，以及国外大模型未使用大量中文数据进行训练，导致模型在错误识别方面表现良好，但在错误类型识别和错误类型定位方面存在局限性。

如表2所示，国内大模型（包括ChatGLM-4、ERNIE-4）在语法错误识别子任务上的F1值分别为67.64和71.33，在错误类型识别子任务上的F1值分别为43.06和51.9，在错误类型定位子任务上的F1值分别为14.54和21.96。这些结果表明，在低资源下，国内大模型在中文语法检测任务上的表现已经接近部分有监督模型，可以有效的完成中文语法检测任务。然而，这些模型普遍存在准确率相对较低而召回率较高的情况，因此可以在模型输出结果基础上做进一步改进。

| 模型 | P | R | F0.5 |
|---|---|---|---|
| 序列到编辑 | | | |
| HRG (Hinson et al., 2020) | 36.79 | 27.82 | 34.56 |
| SG-GEC (Wu and Wu, 2022) | 50.56 | 25.24 | 42.11 |
| 序列到序列 | | | |
| AliGM(Zhou et al., 2018a) | 41 | 13.75 | 29.36 |
| YouDao(Fu et al., 2018a) | 35.24 | 18.64 | 29.91 |
| BLCU(Li et al., 2019) | 47.63 | 12.56 | 30.57 |
| Qiu and Qu (2019) | 36.88 | 18.94 | 31.01 |
| MaskGEC(Zhao and Wang, 2020) | 44.36 | 22.18 | 36.97 |
| TEA(王辰成et al., 2020) | 39.43 | 22.8 | 34.41 |
| WCDA(Tang et al., 2021) | 47.29 | 23.89 | 39.49 |
| Copy(Zhao et al., 2019) | 51.25 | 32.55 | **45.97** |
| SynGEC(Zhang et al., 2022b) | 49.96 | **33.04** | 45.32 |
| TemplateGEC (Li et al., 2023a) | **54.50** | 27.40 | 45.50 |
| 国外大模型 | | | |
| LLaMa-2-7B-Chat | 11.79 | 11.46 | 11.72 |
| GPT-3.5-Turbo | 26.64 | 22.75 | 25.76 |
| 国内大模型 | | | |
| Qwen-7B-Chat | 19.95 | 22.53 | 20.42 |
| BaiChuan-7B-Chat | 20.87 | 23.28 | 21.31 |
| ChatGLM-3-6B | 19.52 | 21.09 | 19.82 |
| ChatGLM-4 | 24.27 | 32.34 | 25.72 |
| ERNIE-4 | **32.78** | **38.15** | **33.72** |

表 3: 大模型在NLPCC2018测试集上结果

## 4.2 中文语法纠错实验结果

如表3所示，大模型在中文语法纠错任务上的表现与在中文语法检测任务相似，均不如现有的有监督模型，但在低资源下，仍展现出可观的性能。此外，在中文语法纠错任务上，国外大模型的表现不及国内大模型，且参数规模较小的大模型（6-7B）的性能也低于大型参数模型。

6-7B参数量级的国内大模型（包括ChatGLM-3-6B、Qwen-7B-Chat和BaiChuan-2-7B-Chat）在中文语法纠错任务上的F0.5值分别为19.82，20.42和21.31，表现较弱。这种性能不足可能由两个主要因素导致：首先，中文语法纠错任务相较于中文语法检测任务更为复杂，它不仅要求模型能够识别错误，还要求模型能够进行错误的修正，这涉及到词序、标点使用和词汇选择等多个错误类型，增加了任务的难度。其次，这些模型可能未经过针对中文语法纠错数据集的专门训练，且模型参数规模相对较小，导致模型在学习过程中可能未能充分捕捉到任务所需的特征和规律。

国外大模型（如GPT-3.5-Turbo和LLaMa-2-7B-Chat）在中文语法纠错上的F0.5值分别为11.72和25.76，表现不佳，这可能是因为它们在训练时未接触大量中文数据，无法充分适应中文的语言特性。例如，源输入“中国，悠久的历史，灿烂的文化，真是在历史上最难忘的国家。”，参考答案为“中国，拥有悠久的历史，灿烂的文化，真是在历史上让人最难忘的国


(c) 2024 中国中文信息学会计算语言学专业委员会

家。”，GPT-3.5-Turbo模型的输出结果为“中国，悠久的历史，灿烂的文化，真是历史上最难忘的国家。”。这表明模型的纠错结果更倾向于英文的表达方式。

国内大模型在中文语法纠错任务上的表现较为出色，ChatGLM-4和ERNIE-4的F0.5值分别为25.72和33.72。虽然与现有的有监督模型相比存在一定差距，但在低资源下，已经展现出卓越的性能。如表3所示，国内大模型普遍存在召回率高而准确率低的情况。这一现象可能是由于模型主要接受了大量正确句子的训练，而在面对错误句子时，模型倾向于根据其训练数据对其进行修正。由于同一语意的句子可以有多种表达方式，这种修正方法可能不适合正式或非标准的语言环境，有时甚至可能导致过度校正。此外，模型在语义理解方面也存在不足，具体例子如表4所示。表4的标注旨在对比模型输入、模型输出与正确答案之间的差异，与传统的纠错标注方法有所区别。

| 类型 | 输入 | 模型输出 | 正确答案 |
|---|---|---|---|
| 标点符号使用不当 | 因为读书准备就业的时间越少了。 | 因为读书准备就业的时间**更少**了。 | 因为读书，准备就业的时间**更少**了。 |
| 拼写错误 | **我看来**北京人很爽快。 | **在我看来**，北京人很爽快。 | **在我看来**北京人很爽快。 |
| | 人生就是过山车，**又高又低**，**又**光明又黑暗。 | 人生就是过山车，**又高有低**，**有光明**又有黑暗。 | 人生就是过山车，**有高有低**，**有光明**又有黑暗。 |
| 搭配不当 | 不**带**口罩的有些朋友们特别后悔**了**。 | 有些不**带**口罩的朋友们特别后悔。 | 有些不**戴**口罩的朋友们特别后悔。 |
| | 现代社会的就业**难太厉害**。 | 现代社会的就业**太难厉害**。 | 现代社会的就业**太难**。 |
| 量词使用不当 | 但是这种想法太**近视眼的**，而有很大的错误。 | 但是这种想法太**近视**，而且有很大的错误。 | 但是这种想法太**短浅**，而且有很大的错误。 |
| | 本人在贵公司实习了**二个**多月，**此时间**经验了许多新鲜的事情。 | 本人在贵公司实习了**二个**多月，**此间**经历了许多新鲜的事情。 | 本人在贵公司实习了**两个**多月，**此间**经历了许多新鲜的事情。 |
| 名词使用不当 | 十二月是瑞典一年中黑夜最长白天短的**一天**。 | 十二月是瑞典一年中黑夜最长白天最短的**一个月**。 | 十二月是瑞典一年中黑夜最长白天最短的**一月**。 |
| | 勤奋刻苦，不折不挠的精神是从黑暗走向光明的人的必要条件。 | 勤奋刻苦，不折不挠的精神是从黑暗走向光明的必要条件。 | 勤奋刻苦，不折不挠的精神是**人**从黑暗走向光明的必要条件。 |
| 过度校正 | 四个人在**同桌**吃一顿饭是很难得。 | 四个人在**桌**吃一顿饭是很难得的。 | 四个人在**同桌**吃一顿饭是很难得的。 |
| | 这样，你就会尝到泰国人**死爱**的味道。 | 你将会尝到泰国人**喜爱**的味道 | 这样，你就会尝到泰国人**爱死**的味道。 |
| | 抽烟对身体非常不好，对身体**非常危害**。 | 抽烟对身体非常不好，对身体**危害极大**。 | 抽烟对身体非常不好，对身体**危害非常大**。 |

表 4: ERNIE-4在NLPCC2018测试集中个别样例

## 5 大模型提示策略影响分析

本节将从三个方面探讨影响ERNIE-4模型在中文语法纠错任务中性能的因素：提示模板中的关键字选择、提示策略的设计以及少样本思维链提示策略中样本的数量。

### 5.1 提示模板中的关键字对大模型性能的影响

鉴于大模型提示模板中的关键词对其性能具有显著影响，本文在 3.4提出的None模板基础上，设计了多种不同的关键词提示，并通过调用ERNIE-4接口进行了测试。这些关键词分为两大类：“直接指示”和“角色扮演”。在“直接指示”类别中，模板明确指示模型执行语法纠错的指令，例如：“查找并校对输入句子中的任何语法错误。”、“查找并修复这个句子中的任何错误。”以及“在不改变原语义和原语言的情况下，润色输入句子。”。这些模板旨在引导模型专注于纠正语法错误，同时保持句子的原始意义和语言风格。在“角色扮演”类别中，模板通过指定模型扮演与语法纠错相关的特定角色来进行提示，例如：“作为专业的中文语法纠错工具”、“担任资深的编辑”或“作为中文教师”。这些模板旨在引导模型从特定专家的角度出发，执行语法纠错任务。具体的提示模板示例和详细信息见附录 8。

如附录 8所示，测试的几种提示模板均能够实现语法纠错的任务。然而，部分模板倾向于在源句的基础上进行不必要的扩展，或是改变原有的句式结构，从而导致与正确答案相比改动


(c) 2024 中国中文信息学会计算语言学专业委员会

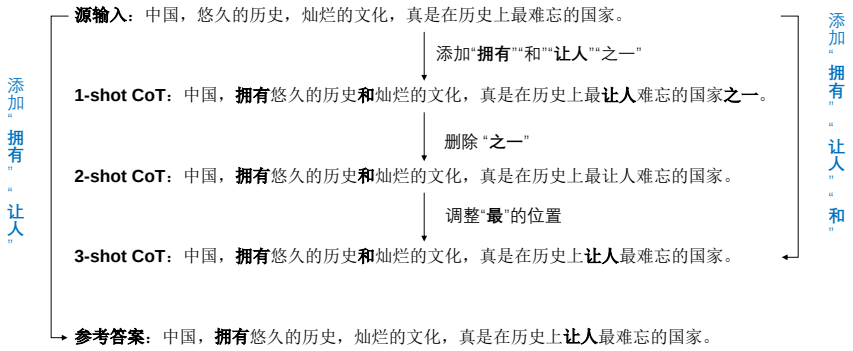



图 1: ERNIE-4使用Few sot-CoT提示策略的输出样例

较大，出现过度校正的现象。因此，本文选取了修改正确且改动最小的关键字作为基础模板的关键字，即“你是一个专业的中文语法纠错工具，具备准确识别和纠正语法错误的能力。”。

### 5.2 提示策略对大模型性能的影响

| ERNIE-4 | NLPCC2018测试集 | | |
|---|---|---|---|
| | P | R | F0.5 |
| None | 30.37 | 39.59 | 31.86 |
| Explain | 30.77 | **40.19** | 32.28 |
| CoT | 32.47 | 37.36 | 32.63 |
| Few-shot CoT | **32.78** | 38.15 | **33.72** |

表 5: 不同提示策略对中文语法纠错任务性能的影响

如表5所示，不同提示模板对ERNIE-4模型在中文语法纠错任务中的性能表现产生了显著影响。性能表现从低到高依次为：无额外解释的提示模板（None）、加入专家解释的提示模板（Explain）、加入思维链（CoT）的提示模板以及加入纠错示例的提示模板（Few-shot CoT）。

1）当大模型仅接收到执行语法纠错任务的指令而缺乏额外解释时，模型必须依赖其先前学习的知识来推断语法纠错的本质，这种理解可能与语法纠错的真实意图存在偏差，导致性能最差。2）当模型被告知语法纠错任务的具体定义时，能够依据这一明确的指示完成任务，从而减少了理解上的偏差，模型性能因此得到提升。3）如果在提示模板中明确指出语法纠错的步骤，即构建思维链，这将引导大模型模型按照标准化的推理过程来执行任务，从而更有效地提高模型性能。4）在思维链的基础上，提供额外的纠错示例能够使模型学习到示例中蕴含的丰富信息，从而更深入地理解语法纠错任务，并实现更卓越的性能表现。

### 5.3 少样本思维链提示策略中样本数量对大模型性能的影响

| ERNIE-4 | NLPCC2018测试集 | | |
|---|---|---|---|
| | P | R | F0.5 |
| 1-shot CoT | 31.05 | 37.36 | 32.14 |
| 2-shot CoT | 32.20 | **38.94** | 33.35 |
| 3-shot CoT | **32.78** | 38.15 | **33.72** |

表 6: 大模型提示模板中样本数量对中文语法纠错性能的影响

为研究少量示例提示模板对大模型纠错能力的影响，本文通过调整提示模板中示例的数量进行了实验。如表6所示，随着提示模板中纠错示例数量的增加，大模型的纠错准确率显著提高，尤其在处理复杂和微妙的语言错误时效果显著。如图 1所示。这可能是因为当大模型接收到包含示例的提示模板时，它们能够更快地识别文本中的不一致性，并从中归纳出纠错的规则。这种学习机制不仅使大模型能够机械地纠正错误，而且能够更深入地理解文本纠错的本


(c) 2024 中国中文信息学会计算语言学专业委员会

质，即如何在保持原文意蕴不变的前提下，提高语言的准确性和流畅性。实验结果验证了在提示模板中添加少量示例可以提升大模型纠错能力方面的有效性，通过模仿和学习大量示例，大模型能够增强执行语法纠错任务的能力。

| | W | S | M | R | 无 |
|---|---|---|---|---|---|
| gold | **202** | 1961 | 963 | 698 | 17 |
| None | 162(25%) | 2108(22%) | 1117(13%) | 802(21%) | 206 |
| Explain | 177(26%) | 2092(22%) | 1079(14%) | 873(21%) | 206 |
| CoT | 170(25%) | 2205(21%) | 1119(14%) | 895(20%) | 210 |
| 1-shot CoT | 157(28%) | 2398(17%) | 1141(12%) | 907(18%) | **277** |
| 2-shot CoT | 169(24%) | 2322(20%) | **1165(14%)** | **921(19%)** | 200 |
| 3-shot CoT | 168(24%) | **2401(17%)** | 1140(11%) | 919(19%) | 263 |

表 7: 基础错误类型分析（数量/准确率）

## 6 针对不同错误类型的纠错结果分析

为了深入研究大模型在中文语法纠错任务中的性能表现，本文对ERNIE-4在不同错误类型上的纠错效果进行了细致分析。本文采用了ChERRANT工具(Zhang et al., 2022a)来提取错误类型，这些错误类型被划分为基础错误和详细错误两大类。具体来说，本文通过对比大模型生成结果与标准答案之间的差异，对大模型在不同错误类型上的表现进行了评估。基础错误和详细错误的分析结果分别展示在表7和8中，其中，gold代表标准答案，而None、Explain、CoT、1-shot CoT、2-shot CoT和3-shot CoT则代表了大模型所采用的不同提示模板。

| | W | PUNCT | SPELL | NUM | PREP | OTHER | ADV |
|---|---|---|---|---|---|---|---|
| gold | **202** | 323 | **390** | 34 | **230** | 639 | 226 |
| None | 162(25%) | 491(11%) | 314(54%) | 39(17%) | 194(24%) | 631(8%) | 242(16%) |
| Explain | 177(26%) | 505(11%) | 319(52%) | 46(21%) | 206(24%) | 648(8%) | 257(13%) |
| CoT | 170(25%) | 515(12%) | 317(55%) | 47(17%) | 213(24%) | 722(6%) | 260(14%) |
| 1-shot CoT | 157(28%) | 511(10%) | 296(49%) | **51(19%)** | 221(22%) | 758(5%) | 293(11%) |
| 2-shot CoT | 169(24%) | **520(11%)** | 314(52%) | 47(19%) | 219(25%) | 755(6%) | 283(13%) |
| 3-shot CoT | 168(24%) | 478(10%) | 287(52%) | 45(24%) | 206(25%) | **792(6%)** | **297(11%)** |
| | AUX | NOUN | VERB | CONJ | PRON | ADJ | QUAN |
| gold | 286 | 369 | 735 | 118 | 141 | 107 | 24 |
| None | 467(17%) | 391(20%) | 787(23%) | 163(11%) | 185(12%) | 97(18%) | 26(38%) |
| Explain | 448(19%) | 395(19%) | 765(24%) | 151(12%) | 173(15%) | 101(16%) | 30(23%) |
| CoT | 466(18%) | 422(19%) | 790(22%) | 154(14%) | 185(14%) | 99(18%) | 29(27%) |
| 1-shot CoT | 481(16%) | 449(15%) | **864(18%)** | 196(8%) | 189(12%) | 106(16%) | 31(22%) |
| 2-shot CoT | **498(17%)** | 432(20%) | 860(21%) | 177(10%) | 179(15%) | 96(13%) | 28(25%) |
| 3-shot CoT | 479(16%) | **450(15%)** | 885(17%) | **190(9%)** | **198(12%)** | **120(17%)** | **33(21%)** |

表 8: 详细错误类型分析（数量/准确率）

如表7所示，W代表词序错误，S代表替换错误，M代表缺失错误，R代表冗余错误。在基础错误类型上，从准确率的角度来看，大模型在处理缺失错误方面的表现较弱，而在词序错误方面则表现较为出色。从识别出的错误数量上来分析，大模型在纠错过程中不倾向于修改词序错误，而更倾向于修正替换错误。总体来看，虽然生成的替换错误数量较多，但正确率并不高，这表明大模型倾向于将正确的词语修改为模型先前学习到的词语，从而导致过度校正的问题。相比之下，生成的词序错误数量较少，但正确率较高，这可能是因为大模型的训练数据偏向于口语化表达，而测试集中的数据偏向于学术化表达，导致大模型在面对测试集中口语化句子时，倾向于认为其是正确的。

如表8所示，其中SPELL代表拼写错误，PREP代表介词错误，NUM代表数字错误，ADJ代表形容词错误，QUAN代表量词错误等。在详细错误类型方面，大模型在识别SPELm56778L和PREP错误类型上的数量低于gold，而在其他错误类型中，大模型识别出的错误数量均超过了gold。从准确率的角度来看，大模型在处理PUNCT、CONJ和PRON错误类型时表现较差，而在SPELL和QUAN错误类型上表现较好。这说明大模型在纠错过程中不倾


(c) 2024 中国中文信息学会计算语言学专业委员会

向于修改拼写错误，但其正确率较高，这可能是因为大模型在大量数据训练中更多地学习了正确的token，而未能学习到测试集中的拼写错误token。大模型在量词错误上的良好表现可能是因为中文语法中量词错误相对容易修正。另一方面，大模型在介词错误上的识别数量少于标准答案，这可能因为中文语法中介词用法的复杂性，导致大模型未能充分学习中文介词的使用规则。这些分析揭示了大模型在中文语法纠错任务中的优势和局限性，为未来针对特定错误类型优化大模型的性能提供了一些参考依据。

## 7 总结

本文采用不同提示策略，评估了国内外主流大模型在中文语法检测和中文语法纠错任务中的性能表现，并对大模型的纠错能力进行了基于错误类型的细粒度评价。实验结果表明，国内大模型在中文文本纠错任务上的表现优于国外大模型，尤其是采用少样本思维链提示策略时，性能表现最为显著，对词序错误和拼写错误的纠正准确率较高。尽管大模型在低资源下能够有效完成中文文本纠错任务，但与有监督模型相比，仍存在一定差距，并存在过度校正的问题。

为了进一步优化大模型的性能，未来的研究可以从以下三个方面着手：1）对大模型的纠错结果进行再训练，以减少过度校正的问题；2）将大模型与有监督的小模型相结合，通过知识蒸馏的方式，在大模型中提取有针对性的知识，以提高小模型的纠错性能；3）通过强化或改进少样本思维链提示策略，进一步提升大模型在中文文本纠错任务中的表现。

## 致谢

感谢审稿人帮助改善论文的建议。本文受国家自然科学基金青年项目（62306284），中国博士后科学基金面上项目（2023M743189），河南省自然科学基金青年项目（232300421386）资助。

## 参考文献

Abhijeet Awasthi, Sunita Sarawagi, Rasna Goyal, Sabyasachi Ghosh, and Vihari Piratla. 2019. Parallel iterative edit models for local sequence transduction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4260–4270, Hong Kong, China, November. Association for Computational Linguistics.

Hongxiang Chang, Yang Liu, Meng Xu, Yingying Wang, Cunliang Kong, Liner Yang, Yang Erhong, Maosong Sun, Gaoqi Rao, Renfen Hu, Zhenghao Liu, 鸿翔常, 洋刘, 萌徐, 莹莹王, 存良孔, 麟儿杨, 尔弘杨, 茂松孙, 高琦饶, 韧奋胡, and 正皓刘. 2023. CCL23-eval 任务7总结报告: 汉语学习者文本纠错(overview of CCL23-eval task: Chinese learner text correction). In Maosong Sun, Bing Qin, Xipeng Qiu, Jing Jiang, and Xianpei Han, editors, *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 3: Evaluations)*, pages 239–249, Harbin, China, August. Chinese Information Processing Society of China.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey on in-context learning.

Yaxin Fan, Feng Jiang, Peifeng Li, and Haizhou Li. 2023. Grammargpt: Exploring open-source llms fornative chinese grammatical error correction withsupervised fine-tuning. In Fei Liu, Nan Duan, Qingting Xu, and Yu Hong, editors, *Natural Language Processing and Chinese Computing*, pages 69–80, Cham. Springer Nature Switzerland.

Tao Fang, Jinpeng Hu, Derek F. Wong, Xiang Wan, Lidia S. Chao, and Tsung-Hui Chang. 2023a. Improving grammatical error correction with multimodal feature integration. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9328–9344, Toronto, Canada, July. Association for Computational Linguistics.

Tao Fang, Yang Shu, Kaixin Lan, Derek Wong, Jinpeng Hu, Lidia Chao, and Yue Zhang. 2023b. Is chatgpt a highly fluent grammatical error correction system? a comprehensive evaluation, 04.

Tao Fang, Shu Yang, Kaixin Lan, Derek F. Wong, Jinpeng Hu, Lidia S. Chao, and Yue Zhang. 2023c. Is chatgpt a highly fluent grammatical error correction system? a comprehensive evaluation. *ArXiv*, abs/2304.01746.


(c) 2024 中国中文信息学会计算语言学专业委员会

Yuejiao Fei, Leyang Cui, Sen Yang, Wai Lam, Zhenzhong Lan, and Shuming Shi. 2023. Enhancing grammatical error correction systems with explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7489–7501, Toronto, Canada, July. Association for Computational Linguistics.

Kai Fu, Jin Huang, and Yitao Duan. 2018a. Youdao's winning solution to the nlpcc-2018 task 2 challenge: A neural machine translation approach to chinese grammatical error correction. In *Natural Language Processing and Chinese Computing*.

Ruiji Fu, Zhengqi Pei, Jiefu Gong, Wei Song, Dechuan Teng, Wanxiang Che, Shijin Wang, Guoping Hu, and Ting Liu. 2018b. Chinese grammatical error diagnosis using statistical and prior knowledge driven features with probabilistic ensemble enhancement. In Yuen-Hsien Tseng, Hsin-Hsi Chen, Vincent Ng, and Mamoru Komachi, editors, *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 52–59, Melbourne, Australia, July. Association for Computational Linguistics.

Yingjie Han, Yingjie Yan, Yangchao Han, Rui Chao, and Hongying Zan. 2020. Chinese grammatical error diagnosis based on RoBERTa-BiLSTM-CRF model. In Erhong YANG, Endong XUN, Baolin ZHANG, and Gaoqi RAO, editors, *Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 97–101, Suzhou, China, December. Association for Computational Linguistics.

Charles Hinson, Hen-Hsen Huang, and Hsin-Hsi Chen. 2020. Heterogeneous recycle generation for Chinese grammatical error correction. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2191–2201, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada, July. Association for Computational Linguistics.

Masahiro Kaneko and Naoaki Okazaki. 2023. Controlled generation with prompt insertion for natural language explanations in grammatical error correction. *ArXiv*, abs/2309.11439.

Masahiro Kaneko, Kengo Hotate, Satoru Katsumata, and Mamoru Komachi. 2019. TMU transformer system using BERT for re-ranking at BEA 2019 grammatical error correction on restricted track. In Helen Yannakoudakis, Ekaterina Kochmar, Claudia Leacock, Nitin Madnani, Ildikó Pilán, and Torsten Zesch, editors, *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 207–212, Florence, Italy, August. Association for Computational Linguistics.

Masahiro Kaneko, Sho Takase, Ayana Niwa, and Naoaki Okazaki. 2022. Interpretability for language learners using example-based grammatical error correction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7176–7187, Dublin, Ireland, May. Association for Computational Linguistics.

Shun Kiyono, Jun Suzuki, Masato Mita, Tomoya Mizumoto, and Kentaro Inui. 2019. An empirical study of incorporating pseudo data into grammatical error correction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1236–1242, Hong Kong, China, November. Association for Computational Linguistics.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Sang Kwon, Gagan Bhatia, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. Beyond English: Evaluating LLMs for Arabic grammatical error correction. In Hassan Sawaf, Samhaa El-Beltagy, Wajdi Zaghouani, Walid Magdy, Ahmed Abdelali, Nadi Tomeh, Ibrahim Abu Farha, Nizar Habash, Salam Khalifa, Amr Keleg, Hatem Haddad, Imed Zitouni, Khalil Mrini, and Rawan Almatham, editors, *Proceedings of ArabicNLP 2023*, pages 101–119, Singapore (Hybrid), December. Association for Computational Linguistics.


(c) 2024 中国中文信息学会计算语言学专业委员会

Changliang Li and Ji Qi. 2018. Chinese grammatical error diagnosis based on policy gradient LSTM model. In Yuen-Hsien Tseng, Hsin-Hsi Chen, Vincent Ng, and Mamoru Komachi, editors, *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 77–82, Melbourne, Australia, July. Association for Computational Linguistics.

Si Li, Jianbo Zhao, Guirong Shi, Yuanpeng Tan, Huifang Xu, Guang Chen, Haibo Lan, and Zhiqing Lin. 2019. Chinese grammatical error correction based on convolutional sequence to sequence model. volume 7, pages 72905–72913.

Bei Li, Quan Du, Tao Zhou, Yi Jing, Shuhan Zhou, Xin Zeng, Tong Xiao, JingBo Zhu, Xuebo Liu, and Min Zhang. 2022a. ODE transformer: An ordinary differential equation-inspired model for sequence generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8335–8351, Dublin, Ireland, May. Association for Computational Linguistics.

Jiquan Li, Junliang Guo, Yongxin Zhu, Xin Sheng, Deqiang Jiang, Bo Ren, and Linli Xu. 2022b. Sequence-to-action: Grammatical error correction with action guided sequence generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10974–10982, Jun.

Yinghao Li, Xuebo Liu, Shuo Wang, Peiyuan Gong, Derek F. Wong, Yang Gao, Heyan Huang, and Min Zhang. 2023a. TemplateGEC: Improving grammatical error correction with detection template. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6878–6892, Toronto, Canada, July. Association for Computational Linguistics.

Yinghui Li, Haojing Huang, Shirong Ma, Yong Jiang, Y. Li, F. Zhou, Haitao Zheng, and Qingyu Zhou. 2023b. On the (in)effectiveness of large language models for chinese text correction. *ArXiv*, abs/2307.09007.

Deng Liang, Chen Zheng, Lei Guo, Xin Cui, Xiuzhang Xiong, Hengqiao Rong, and Jinpeng Dong. 2020. BERT enhanced neural machine translation and sequence tagging model for Chinese grammatical error diagnosis. In Erhong YANG, Endong XUN, Baolin ZHANG, and Gaoqi RAO, editors, *Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 57–66, Suzhou, China, December. Association for Computational Linguistics.

Xuebo Liu, Longyue Wang, Derek F. Wong, Liang Ding, Lidia S. Chao, and Zhaopeng Tu. 2021. Understanding and improving encoder layer fusion in sequence-to-sequence learning.

Mengsay Loem, Masahiro Kaneko, Sho Takase, and Naoaki Okazaki. 2023. Exploring effectiveness of GPT-3 in grammatical error correction: A study on performance and controllability in prompt-based methods. In Ekaterina Kochmar, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Nitin Madnani, Anaïs Tack, Victoria Yaneva, Zheng Yuan, and Torsten Zesch, editors, *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 205–219, Toronto, Canada, July. Association for Computational Linguistics.

Yikang Luo, Zuyi Bao, Chen Li, and Rui Wang. 2020. Chinese grammatical error diagnosis with graph convolution network and multi-task learning. In Erhong YANG, Endong XUN, Baolin ZHANG, and Gaoqi RAO, editors, *Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 44–48, Suzhou, China, December. Association for Computational Linguistics.

Eric Malmi, Sebastian Krause, Sascha Rothe, Daniil Mirylenka, and Aliaksei Severyn. 2019. Encode, tag, realize: High-precision text editing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5054–5065, Hong Kong, China, November. Association for Computational Linguistics.

Masato Mita and Hitomi Yanaka. 2021. Do grammatical error correction models realize grammatical generalization? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4554–4561, Online, August. Association for Computational Linguistics.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. 2020. GECToR – grammatical error correction: Tag, not rewrite. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–170, Seattle, WA, USA → Online, July. Association for Computational Linguistics.


(c) 2024 中国中文信息学会计算语言学专业委员会

Kevin Parnow, Zuchao Li, and Hai Zhao. 2021. Grammatical error correction as GAN-like sequence labeling. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3284–3290, Online, August. Association for Computational Linguistics.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2023. Reasoning with language model prompting: A survey. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5368–5393, Toronto, Canada, July. Association for Computational Linguistics.

Zhaoquan Qiu and Youli Qu. 2019. A two-stage model for chinese grammatical error correction. *IEEE Access*, 7:146772–146777.

Gaoqi Rao, Qi Gong, Baolin Zhang, and Endong Xun. 2018. Overview of NLPTEA-2018 share task Chinese grammatical error diagnosis. In Yuen-Hsien Tseng, Hsin-Hsi Chen, Vincent Ng, and Mamoru Komachi, editors, *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 42–51, Melbourne, Australia, July. Association for Computational Linguistics.

Hongkai Ren, Liner Yang, and Endong Xun. 2018. A sequence to sequence learning for chinese grammatical error correction. In Min Zhang, Vincent Ng, Dongyan Zhao, Sujian Li, and Hongying Zan, editors, *Natural Language Processing and Chinese Computing*, pages 401–410, Cham. Springer International Publishing.

Murray Shanahan. 2024. Talking about large language models. *Commun. ACM*, 67(2):68–79, jan.

Yixiao Song, Kalpesh Krishna, Rajesh Bhatt, Kevin Gimpel, and Mohit Iyyer. 2023. Gee! grammar error explanation with large language models. *ArXiv*, abs/2311.09517.

Felix Stahlberg and Shankar Kumar. 2020. Seq2Edits: Sequence transduction using span-level edit operations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5147–5159, Online, November. Association for Computational Linguistics.

Xin Sun and Houfeng Wang. 2022. Adjusting the precision-recall trade-off with align-and-predict decoding for grammatical error correction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 686–693, Dublin, Ireland, May. Association for Computational Linguistics.

Xin Sun, Tao Ge, Furu Wei, and Houfeng Wang. 2021. Instantaneous grammatical error correction with shallow aggressive decoding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5937–5947, Online, August. Association for Computational Linguistics.

Minghuan Tan, Min Yang, and Ruifeng Xu. 2023. Focal training and tagger decouple for grammatical error correction. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5978–5985, Toronto, Canada, July. Association for Computational Linguistics.

Zecheng Tang, Yixin Ji, Yibo Zhao, and Junhui Li. 2021. 基于字词粒度噪声数据增强的中文语法纠错(Chinese grammatical error correction enhanced by data augmentation from word and character levels). In Sheng Li, Maosong Sun, Yang Liu, Hua Wu, Kang Liu, Wanxiang Che, Shizhu He, and Gaoqi Rao, editors, *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 813–824, Huhhot, China, August. Chinese Information Processing Society of China.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Yu Wang, Yuelin Wang, Kai Dang, Jie Liu, and Zhuo Liu. 2021. A comprehensive survey of grammatical error correction. *ACM Trans. Intell. Syst. Technol.*, 12(5), dec.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Huai hsin Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903.


(c) 2024 中国中文信息学会计算语言学专业委员会

Xiuyu Wu and Yunfang Wu. 2022. From spelling to grammar: A new framework for Chinese grammatical error correction. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 889–902, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Haihua Xie, Zhiyou Chen, Jing Cheng, Xiaoqing Lyu, and Zhi Tang. 2020. 基于数据增强和多任务特征学习的中文语法错误检测方法(Chinese grammar error detection based on data enhancement and multi-task feature learning). In Maosong Sun, Sujian Li, Yue Zhang, and Yang Liu, editors, *Proceedings of the 19th Chinese National Conference on Computational Linguistics*, pages 761–770, Haikou, China, October. Chinese Information Processing Society of China.

Benfeng Xu, An Yang, Junyang Lin, Quang Wang, Chang Zhou, Yongdong Zhang, and Zhendong Mao. 2023. Expertprompting: Instructing large language models to be distinguished experts. *ArXiv*, abs/2305.14688.

Konstantin Yakovlev, Alexander Podolskiy, Andrey Bout, Sergey Nikolenko, and Irina Piontkovskaya. 2023. GEC-DePenD: Non-autoregressive grammatical error correction with decoupled permutation and decoding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1546–1558, Toronto, Canada, July. Association for Computational Linguistics.

Zheng Yuan, Shiva Taslimipoor, Christopher Davis, and Christopher Bryant. 2021. Multi-class grammatical error detection for correction: A tale of two systems. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8722–8736, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Yongwei Zhang, Qinan Hu, Fang Liu, and Yueguo Gu. 2018. CMMC-BDRC solution to the NLP-TEA-2018 Chinese grammatical error diagnosis task. In Yuen-Hsien Tseng, Hsin-Hsi Chen, Vincent Ng, and Mamoru Komachi, editors, *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 180–187, Melbourne, Australia, July. Association for Computational Linguistics.

Yue Zhang, Zhenghua Li, Zuyi Bao, Jiacheng Li, Bo Zhang, Chen Li, Fei Huang, and Min Zhang. 2022a. MuCGEC: a multi-reference multi-source evaluation dataset for Chinese grammatical error correction. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3118–3130, Seattle, United States, July. Association for Computational Linguistics.

Yue Zhang, Bo Zhang, Zhenghua Li, Zuyi Bao, Chen Li, and Min Zhang. 2022b. SynGEC: Syntax-enhanced grammatical error correction with a tailored GEC-oriented parser. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2518–2531, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Ying Zhang, Hidetaka Kamigaito, and Manabu Okumura. 2023. Bidirectional transformer reranker for grammatical error correction. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3801–3825, Toronto, Canada, July. Association for Computational Linguistics.

Zewei Zhao and Houfeng Wang. 2020. Maskgec: Improving neural grammatical error correction via dynamic masking. In *AAAI Conference on Artificial Intelligence*.

Jianbo Zhao, Si Li, and Zhiqing Lin. 2018a. Contextualized character representation for Chinese grammatical error diagnosis. In Yuen-Hsien Tseng, Hsin-Hsi Chen, Vincent Ng, and Mamoru Komachi, editors, *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 172–179, Melbourne, Australia, July. Association for Computational Linguistics.

Yuanyuan Zhao, Nan Jiang, Weiwei Sun, and Xiaojun Wan. 2018b. Overview of the nlpcc 2018 shared task: Grammatical error correction. In Min Zhang, Vincent Ng, Dongyan Zhao, Sujian Li, and Hongying Zan, editors, *Natural Language Processing and Chinese Computing*, pages 439–445, Cham. Springer International Publishing.

Wei Zhao, Liang Wang, Kewei Shen, Ruoyu Jia, and Jingming Liu. 2019. Improving grammatical error correction via pre-training a copy-augmented architecture with unlabeled data. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 156–165, Minneapolis, Minnesota, June. Association for Computational Linguistics.


(c) 2024 中国中文信息学会计算语言学专业委员会

Junpei Zhou, Chen Li, Hengyou Liu, Zuyi Bao, Guangwei Xu, and Linlin Li. 2018a. Chinese grammatical error correction using statistical and neural models. In Min Zhang, Vincent Ng, Dongyan Zhao, Sujian Li, and Hongying Zan, editors, *Natural Language Processing and Chinese Computing*, pages 117–128, Cham. Springer International Publishing.

Junpei Zhou, Chen Li, Hengyou Liu, Zuyi Bao, Guangwei Xu, and Linlin Li, 2018b. *Chinese Grammatical Error Correction Using Statistical and Neural Models: 7th CCF International Conference, NLPCC 2018, Hohhot, China, August 26–30, 2018, Proceedings, Part II*, pages 117–128. 08.

王辰成, 杨麟儿, 王莹莹, 杜永萍, and 杨尔弘. 2020. 基于transformer增强架构的中文语法纠错方法. 中文信息学报, 34(6):106–114.

王厚峰罗文. 2024. 大语言模型评测综述. 中文信息学报, 38(1):1–23.

## 8 附录A.测试提示模板样例

CCL 2024

模板
- 模板1：你是一个专业的中文语法纠错工具，具备准确识别和纠正语法错误的能力。
- 模板2：你是一位资深的编辑，请帮助校对以下内容，完成纠错和润色工作。
- 模板3：你是中文教师，教授外国人学习中文。
- 模板4：在不改变原语义和原语言的情况下，润色输入句子。
- 模板5：查找并校对输入句子中的所有语法错误。
- 模板6：查找并修复这个句子中的所有错误。

输入：任何事情都是各有利弊，众所周知越建立工业越对经济方面有所发展。

正确答案：任何事情都是各有利弊，众所周知**越发展工业，经济方面就越**有所发展。

模板1：任何事物都是各有利弊，众所周知，**越发展工业，对经济发展越**有所**助益**。

模板2：任何事物都**各有利弊**，众所周知，**工业的建立和发展对经济有着显著的推动作用**。

模板3：任何事物都**有其利与弊**，众所周知，**工业的发展对经济有着显著的推动作用**。

模板4：任何事物都**存在其正面和负面效应**，众所周知，**工业的兴起和发展对经济的增长具有显著推动作用**。

模板5：任何事物都**各有利弊**，众所周知，**工业越发展，对经济的推动作用越显著**。

模板6：任何事情都**有利有弊**，众所周知，**工业发展得越充分，对经济发展越有利**。

输入：不官是从口、眼、鼻子进去这样会伤害身体的建康。

正确答案：不**管**是从口、眼**还是**鼻子进去，**都**会伤害身体的**健康**。

模板1：**无论**是从口、眼、鼻子进**入，都**会伤害身体**健康**。

模板2：**无论**是**通过**口**腔**、眼**睛还是**鼻子进**入，这些都可能对身体健康造成伤害。**

模板3：**无论**是**通过**口、眼**还是鼻进入，这些都可能对身体健康造成伤害**。

模板4：**无论**是经口、眼**还是**鼻进**入，这些都可能损害**身体**健康**。

模板5：不**管**是**通过**口、眼、鼻进**入，这些都可能对身体造成伤害**。

模板6：不**管**是**通过**口、眼、鼻进**入，这都**会伤害身体的**健康**。

输入：对一个生名来说空气污染是很危害的问题，对身体不好。

正确答案：对一个生**命**来说空气污染是**有**很**大**危害的问题，对身体不好。

模板1：对**于生命**来说，空气污染是**一个**很**严重**的问题，对身体不好。

模板2：对**于生命而言，**空气污染是**一个非常严重**的问题，**它**对身体**健康极为不利**。

模板3：对**于生命**来说，空气污染是**一个非常严重**的问题，对身体**健康有害**。

模板4：对**于任何生物而言，**空气污染**都**是**一个极具**危害的问题，对身体**健康极为不利**。

模板5：对**于生命**来说，空气污染是**一个**很**严重**的问题，对身体**有害**。

模板6：对**于生命**来说，空气污染是**一个**很**严重**的问题，对身体**有害**。

图 2: 提示模板测试样例

CCL 2024


(c) 2024 中国中文信息学会计算语言学专业委员会

# 面向“以A为B”构式语义场景的汉语框架识别数据集构建

**杨沛渊[1,‡]，苏雪峰[1,2,‡]，李俊材[1,‡]，闫智超[1,‡]，柴清华[3,⋆]，李茹[1,4,⋆,†]**
[1]山西大学 计算机与信息技术学院，山西 太原 030006
[2]山西工程科技职业大学 现代物流学院，山西 晋中 030609
[3]山西大学 外国语学院，山西 太原 030006
[4]山西大学 计算智能与中文信息处理教育部重点实验室，山西 太原 030006
[‡]{971859815，455375251，1251972979，751824801}@qq.com
[⋆]{charles，liru}@sxu.edu.cn

## 摘要

汉语中普遍存在一些语义场景，其语义核心不是以单个词语呈现，而是通过句子中的某个特定结构来表达。然而当前公开发表的数据集中，只有极少数的数据集将这种特定结构作为语义单元进行研究。汉语框架语义知识库是进行汉语深层语义分析与推理的优质资源，目前其激活框架的基本单位均为句中的一个词。本文以汉语框架语义知识库为基础，引入构式语法，使用2020《人民日报》语料库，以“以A为B”构式为例，建立了基于“以A为B”构式的汉语框架识别数据集，包含23849条例句，相应框架141个。本文使用多个汉语框架识别模型及大语言模型在该数据集上进行了实验，并针对传统框架识别模型在以构式为目标词的框架识别任务中由于目标词信息匮乏导致的识别困难问题，提出了基于目标词转化和数据增强的两种方法，使模型准确率达到了88.19%，有效提升了模型挖掘构式蕴含的深层语义信息的能力。

**关键词：** 框架语义学；构式语法；“以A为B”；框架识别

# Dataset for Recognizing Chinese Semantic Frames based on the Semantic Scenario of the *“Yi A Wei B”* Construction

**Peiyuan Yang[1,‡], Xuefeng Su[1,2,‡], Juncai Li[1,‡], Zhichao Yan[1,‡], Qinghua Chai[3,⋆], Ru Li[1,4,⋆,†]**
[1]School of Computer and Information Technology，Shanxi University
[2]School of Modern Logistics, Shanxi Vocational University of Engineering Science and Technology
[3]School of Foreign Languages，Shanxi University
[4]Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education
[‡] {971859815, 455375251, 1251972979, 751824801}@qq.com
[⋆] {charles, liru}@sxu.edu.cn

## Abstract

In Chinese, specific sentence structures often carry semantic meaning, a characteristic underexplored in existing datasets. Leveraging the Chinese Frame Semantic Knowledge Base, we conducted deep semantic analysis, focusing on the structure *“Yi A Wei B”* found in the 2020 *“People’s Daily”* corpus. We created a frame recognition dataset with this structure, comprising 23,849 example sentences and 141 corresponding frames. Experiments featuring multiple frame recognition models revealed recognition difficulties due to insufficient target word information. Two approaches, target word transformation and data enhancement, increased model accuracy to 88.19%, thereby enhancing the model’s ability to extract construction-embedded semantics.

**Keywords:** Frame Semantics , Construction Grammar , *“Yi A Wei B”* , Frame Identification

CCL 2024

## 1 引言

汉语框架网（Chinese FrameNet，CFN）(You and Liu, 2005)是以Fillmore (1976)的框架语义学为理论基础，以汉语真实语料为依据，参照伯克利大学的框架语义网（FrameNet，FN）(Baker et al., 1998)构建的汉语框架语义知识库，包括框架库、句子库和词元库。目前已有1397个框架，涉及18360个词元，同时也对69592条语句进行了标注。

目前汉语框架数据集中激活框架的词元均为单一词汇，然而，在某些句子中，独立的目标词难以充分体现复杂的语义场景。如图1所示，图中的语句基于现有的汉语框架语义知识库，只能将“为”当做目标词，激活【范畴化】框架，但【范畴化】框架并不适配该句子的语义场景。在此句话中，“以...为...”是一个表达语义的整体，表示“网络购物等”代表了“数字消费”。以“以...为...”为整体作为“目标词”激起【代表】框架，更能准确体现该句子的语境及含义。因此，为了丰富框架的语义表达，本文引入了构式语法，增加了以构式为“目标词”的框架语义解析数据，提升了框架语义解析能力，进一步实现对语言的深度理解。

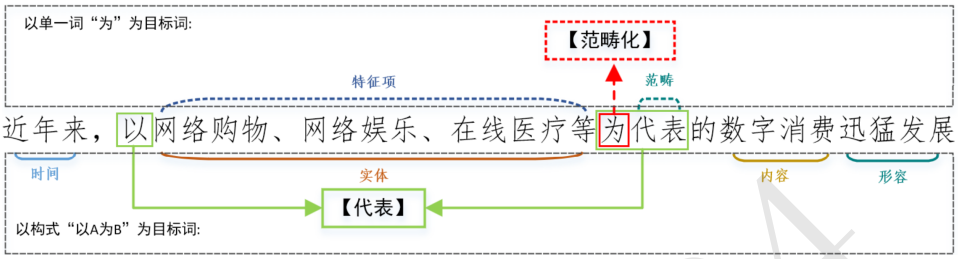



图 1: 框架语义角色标注示例

构式语法(Fillmore et al., 1988)由Fillmore教授于1988年最先提出，其主张语言是由固定的、有意义的单位组成，这些单位被称为构式，既可以是简单的词或短语，也可以是复杂的句子或话语。如，“爱买不买”、“爱说不说”、“爱要不要”对应的构式是“爱V不V”，该构式是一个表达语义的整体，表示对某行为不在意或无所谓。而本文选取的“以A为B”构式是汉语中出现较早的一种结构，它不但在文言文中的使用频率较高，而且在现代汉语中也有着很大的作用(谭世勋, 1985)，还具有一定的代表性和挑战性。一方面，该构式高度抽象，有着非常丰富的子结构，如凝固型构式：以人为本、以食为天等；半凝固型构式：以...为主、以...为例等。另一方面，该构式的语义涵盖范围广泛，能够激活多种框架，如【代表】框架、【方式】框架、【重要性】框架等。再者，该构式在句中能充当各种成分，如主语、宾语、状语等。该构式以其灵活多变的结构和丰富的语义成分，展现出了显著的代表性，是目前框架语义知识库中亟待解决的典型问题之一。

2020年《人民日报》中有新闻报道38540篇，其中有近三分之一篇报道中存在“以A为B”构式。这说明了该构式广泛存在于真实语料中，具有很高的实用价值。此外，在CFN多年的建设中，也存在许多单个目标词激活框架后无法进行语义解析的情况，其中近八分之一的问题语句中存在“以A为B”构式。因此，本文以“以A为B”构式为例，尝试引入构式语法来解决现有汉语框架语义知识库所存在的问题。

本文基于2020《人民日报》语料库，标注了构式及其对应框架的例句23849条，梳理出相应框架141个，构建了超过206万字规模的构式-框架数据集。并且以该数据集为基础，提出了一种面向构式的框架识别方法，并在多种现有模型上进行了实验。实验结果表明，本文提出的方法能有效挖掘构式所蕴含的深层语义信息，显著提高模型针对以“以A为B”构式为目标词的框架识别准确率。

## 2 相关工作

山西大学自2004年开始在英文框架网(FN)的基础上建立汉语框架网(CFN)，基于汉语框架

* 基金项目：国家自然科学基金重点项目（61936012）；山西省科技合作交流专项项目（202204041101016）；山西省重点研发项目（202102020101008）；国家自然科学基金面上项目（62376144）
† 通讯作者 Corresponding Author
©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

语义资源，在框架排歧、零形式识别与填充、篇章结构生成与关系识别等任务上，取得了一系列成果，为汉语深层语义分析与推理提供支撑。如李济洪et al. (2010)等人基于CRF在CFN数据集上进行语义角色标注，屠寒非et al. (2016)等人提出一种基于主动学习的方法并取得一定的效果，Wang et al. (2020)等人提出基于自注意力机制的汉语框架语义角色标注方法以获取句子的长距离信息。2023年山西大学首次公布汉语框架网数据集(CFN2.0)并开放汉语框架语义解析评测(Li et al., 2023a)。目前汉语框架网中框架语义资源包括框架库、词元库、例句库、篇章库等都已具有一定规模，然而其词元库中的目标词均为单一词汇，无法表示构式的深层语义。

构式语法是认知语言学理论体系的重要组成部分。其理论发端于20世纪80年代末Fillmore (1988)等学者对习语的研究，是在反思、批判生成语法及其衍生的其他语法知识模型的基础上产生、发展起来的。该理论一经问世，就受到广泛关注，目前已成为国际语言学界研究的重要内容(文旭and 司卫国, 2021)。

目前的汉语构式研究较多集中在对具体构式的个案研究，且大多集中在构式的语义描写、构式的语用功能特色、构式的认知机制等方面。如孙瑜等人对“先X为敬”构式的生成机制和语用功能的研究，魏在江and 赵帮华 (2024)对隐喻式否定构式的语义生成机制的研究。而这些研究对于汉语构式知识的形式化表示以及在计算机自动分析中的应用等都还很少涉及。

北京大学为了推进构式语法理论在汉语语法领域的研究，并使构式语法理论在计算机领域发挥实际效用，建设了现代汉语构式知识描述数据库(詹卫东and 王佳骏, 2022)，采用类似词库的方式，将真实语料中实际运用的构式形式逐条收录，并详细描写每个构式的内部构成情况、构式整体的语法、语义、语用属性。

本文在北京大学现代汉语构式知识描述数据库基础上，采用其构式界定标准、分类方式和术语体系，以“以A为B”构式为例，将汉语框架网与构式相结合。一方面，丰富了框架的语义表达，解决了部分汉语框架网中存在的不足；另一方面，为构式语法理论在中文信息处理等计算机领域发挥作用提供了新的方案。

## 3 构式-框架语义和数据资源构建

### 3.1 待标注语料筛选

《人民日报》语料库以其规模宏大、主题丰富、规范性强及可靠性高的特点，成为进行各种中文自然语言处理训练的优质选择。如表1所示，本研究选取《人民日报》语句共960418条，根据是否含有构式中的词及是否表达构式含义分为三类。

| 类别 | 样例 | 数量 |
|---|---|---|
| **无“以”和“为”：** | 记者实际探访了独子山矿区。 | 924016 |
| **有“以”和“为”但不是构式：** | ...没有根本利害冲突应该完全可以成为相互需要的合作伙伴。 | 12553 |
| **含有构式：** | 近年来，以网络购物、网络娱乐、在线医疗等为代表的数字消费迅猛发展 | 23084 |

表 1: 人民日报语料分类

确定含有构式中的词（“以”和“为”）后，根据是否表达构式含义分为了正负两类。其中正类是指含有“以A为B”构式的语句,例如: (1)近年来，以网络购物、网络娱乐、在线医疗等为代表的数字消费迅猛发展。(2)今年的政府工作，要以应对国际金融危机、促进经济平稳较快发展为主线。这两句话中的“以网络购物、网络娱乐、在线医疗等为代表”和“以应对国际金融危机、促进经济平稳较快发展为主线” 就是“以A为B”构式，并且句(1)符合“代表”框架所描述的场景；句(2)符合“方式”框架所描述的场景。

相对而言，负类是指含有“以A为B”样式，但不表达“以A为B”构式含义的语句，例如: (1)...没有根本利害冲突应该完全可以成为相互需要的合作伙伴。” (2)天然雪质地松软，并不适合比赛。人造雪可以打造出符合竞技场地技术规范要求的赛道为运动员提供公平竞赛条件。这两句话中都含有“以”和“为”，但是“以”和“为”不单独成词，没有直接的联系，并不属于本文研究的“以A为B”构式。


(c) 2024 中国中文信息学会计算语言学专业委员会

针对区分正负两类数据的问题，本文对抽取出的35637条正负类数据进行了深入的语言学分析，建立了一套符合“以A为B”构式的语言学特征的分类体系。其中主要包含了如下三条规则：

(1)包含的“以”和“为”必须在同一短句中，不允许出现跨标点的情况（顿号除外）。(2)在分词时，“以”和“为”必须被分为单独的词，不允许出现“以”和“为”和其他词粘连的情况。(3)在进行语义依存分析时，“以”和“为”之间必须存在直接关系，且“为”依存于“以”。

根据以上三条规则，本文对含有“以”和“为”的五万余条语料进行了筛选，并选取了其中的1000条数据进行了人工验证。经过测试，根据规则区分出的正负两类数据与人工筛选的吻合率达到了90.6%。证明了所提出规则的有效性，因此，研究依据这套分类体系，从35637条含有“以”和“为”的人民日报数据中选取了23849条正类数据作为下一步框架识别任务中的待标注语料。

## 3.2 数据标注

### 3.2.1 标注规范

汉语框架语义网语义标注的基本单位是由一个框架承担词和若干框架元素组成的语义结构(Li et al., 2024)。框架承担词包括动词、形容词和名词，它们是标注工作的着眼点，称之为目标词。本研究构建的数据集尝试将汉语框架语义网中现有的目标词由单一词汇扩展为构式，即以构式为激活框架的词元。每个标注实例中的目标词（构式）由系统依据位置信息给定，标注人员根据标注实例的语义信息确定待标注实例中目标词的选择是否合理。在确定标注实例的目标词选择无误后，标注人员需要遵循以下三条标注规范进行正式标注：

(1)依据标注实例的语义信息确定目标词边界。句中的“以A为B”构式在“为”之后可能存在修饰限定成分，如“茶马古道位于我国西南地区，以马帮为主要运输方式...”这句话中，“以...为主要运输方式”是一个完整的“以A为B”构式，其中“B”为“主要运输方式”，“主要运输”是对方式的修饰和限定。因此本句中的目标词边界应为“以...为主要运输方式”，而不是“以...为主要”。标注人员在进行标注时需根据语义信息，确定目标词的边界，确保标注数据的准确性。

(2)依据框架含义及框架所对应的框架元素信息确定该句中目标词所匹配的框架。如图2所示，CFN中的框架是一个树状结构，框架之间存在上下位继承关系。例如：【有意行为】框架是一个较为上位的框架，其框架含义为：有意志的生命所进行的活动；【有意创造】框架继承自【有意行为】框架，其框架含义为：创造者可能由成分创造出一个新的实体，创造物；【生产制造】又继承自【有意创造】框架，其框架含义为：生产者出于商业目的，用某种资源生产产品。可以看出，越靠上位的框架，其概念越抽象，适用范围越广；越靠下位的框架，其概念越具体，适用范围越单一。在为待标注实例匹配框架时，会遇到多个框架都匹配的情况。要求标注人员对框架语义知识库中的框架关系有深入了解，根据框架之间的继承关系，遵循“下位框架优先”的原则，选择待标注实例所匹配的框架中框架概念最具体的框架。

(3)在实际标注过程中，有些待标注句子本身存在错误，如构式位置匹配有误或句子长度过长的问题。对于这类句子，标注人员需要根据实际情况，对句子进行适当的调整或删减，以确保标注数据的准确性。

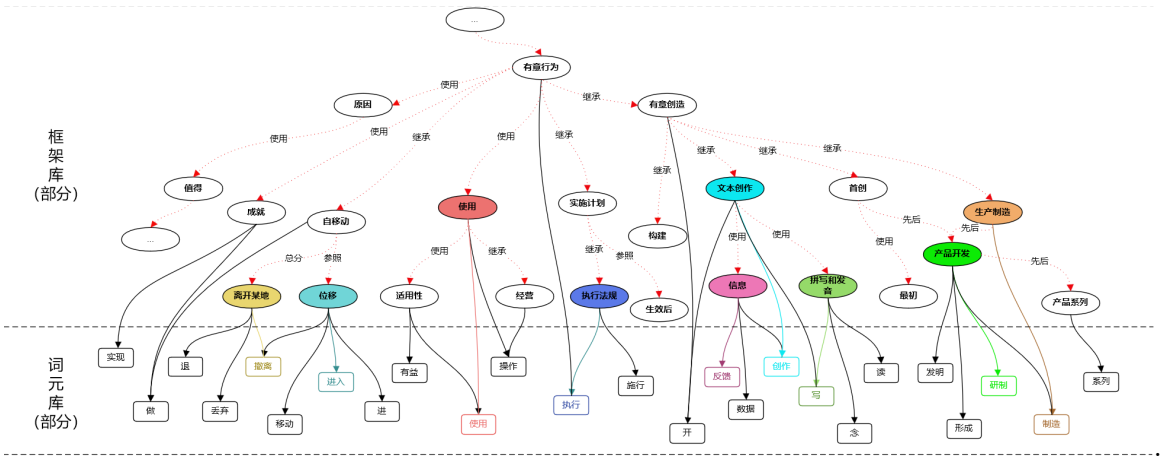



图 2: 框架树


(c) 2024 中国中文信息学会计算语言学专业委员会

### 3.2.2 待标注实例分配

由于“以A为B”构式所表达的含义与“B”有着密切的关系。为了确保人工标注的一致性及准确性，尽量让每个标注人员及审核人员所标注、审核的数据都具有相同或相近的含义。因此在进行人工标注前，根据“B”的词性对“以A为B”构式进行了分类，并将归类后的数据按类别分配于不同的标注人员。如表2所示，分类后确保了每个标注人员分配到的待标注实例均为同一类别。

| “B”类别 | 待标注实例 |
|---|---|
| “B”为名词 | ...要**以**应对国际金融危机、促进经济平稳较快发展**为**主线。<br>**以**规范制度和制约权力**为**核心，针对腐败现象易发多发的领域和环节... |
| “B”为形容词 | 可是，现在的人们不但不复**以**窄窄金莲为美，反异口同韵的诋**为**丑恶。<br>...对什么也有主张，而且**以**扯谎**为**荣。 |
| “B”为动词 | 我们的敌人，哼，只**以**流血**为**享受，而毫无禁忌。<br>这是由心中平静而然，并不是**以**退**为**进。 |

表 2: 构式分类

此外，在分配任务时使用了贪心算法，该算法给每个标注人员和审核人员分配数据时，会尽可能分配标注人员和审核人员曾经标注和审核过的“以A为B”构式，充分的保证了数据标注的准确、高效。

### 3.3 标注流程

本研究基于汉语框架语义知识库及新构建的“以A为B”构式数据集，开展了语料标注工作。整个标注分为两个阶段。第一阶段组织标注人员使用少量数据进行预标注，根据标注结果对标注人员进行讲解及培训。确保标注人员对任务有较深的理解。依据第一阶段的标注经验，对标注流程和系统进行优化，第二阶段展开大规模的标注工作。为了使人工标注更加便捷高效，研究还构建了一个轻量级的标注网站。

参与标注的人员分为标注人员，审核人员及答疑人员，本文将对一个句子的标注称为一个“标注实例”，一个标注实例包含以下基础信息：(1)含有待确定框架的“以A为B”构式完整语句。(2)经过筛选的所有待选择框架。每个标注实例都会被分配给两名不同的标注人员，标注人员根据语义信息及框架含义进行框架选择。在标注人员标注完成后，标注实例进入审核流程。一审人员会对两个标注人员的标注结果提出意见，然后交由二审人员进行最终的审核。二审人员给出最终答案后，经审核的实例结果会返回给标注人员，如果标注人员无异议，则根据最终结果进行学习。如果标注人员有异议则由答疑人员进行解答并给出最终答复。答疑老师会针对标注遇到的问题及时对数据或知识库进行更新或修改。详细标注流程如图3所示。

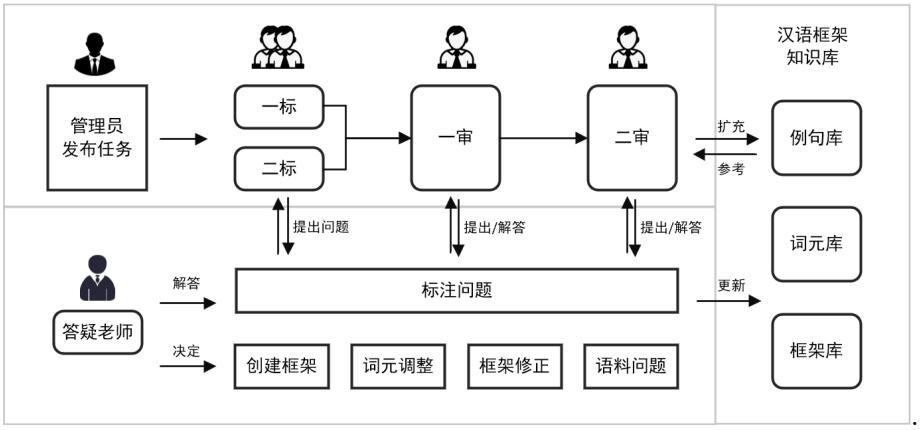



图 3: 标注流程


(c) 2024 中国中文信息学会计算语言学专业委员会

### 3.4 数据集规模及分布

本研究共标注了23849条数据。其中有138个句子长度过长，243个句子中构式的位置未被正确识别，还有384个句子本身存在错误。最终获得了23084条高质量的标注语料，每条语料都被匹配了唯一的框架，该构式-框架数据集共涉及到141个框架，其中131个框架已存在于汉语框架语义知识库中，还有10个框架待创建。数据集中各框架的分布如图4所示。

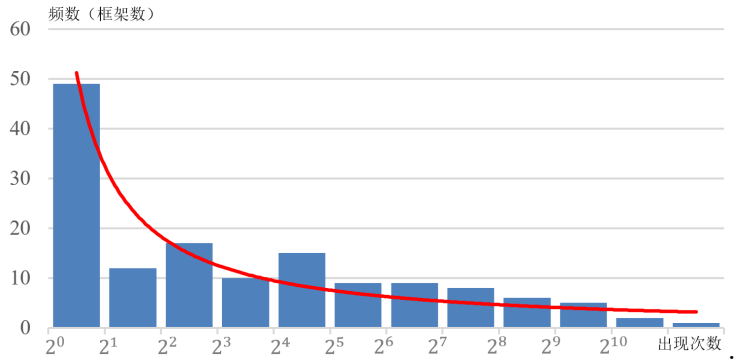



图 4: 框架数量频数分布直方图

由图可见，存在大量框架仅具有少数例句的情况，甚至超过半数的框架仅具有20条以下的例句，而例句数最多的框架则具有2256条例句，虽然呈现长尾分布现象，但符合人类在进行自然语言描述时的现实规律，这种现象增加了数据的复杂性。这些特点对框架语义分析模型提出了较高的要求。

此外，本文对“以A为B”构式进行了梳理，按照“B”的内容对数据集进行了分类，整理出不同类型的“B”内容共计413种。如图5所示，存在一种构式类型对应多种框架的情况，也存在一种框架对应多种构式类型的情况，即构式类型与框架不是简单的一对一的情况。这样的分布情况使得构式-框架数据集具有较高的复杂性，对框架识别模型的识别能力提出了更高的要求。模型不能简单的通过构式的匹配来确定框架，必须学习到构式在句中的深层次含义。

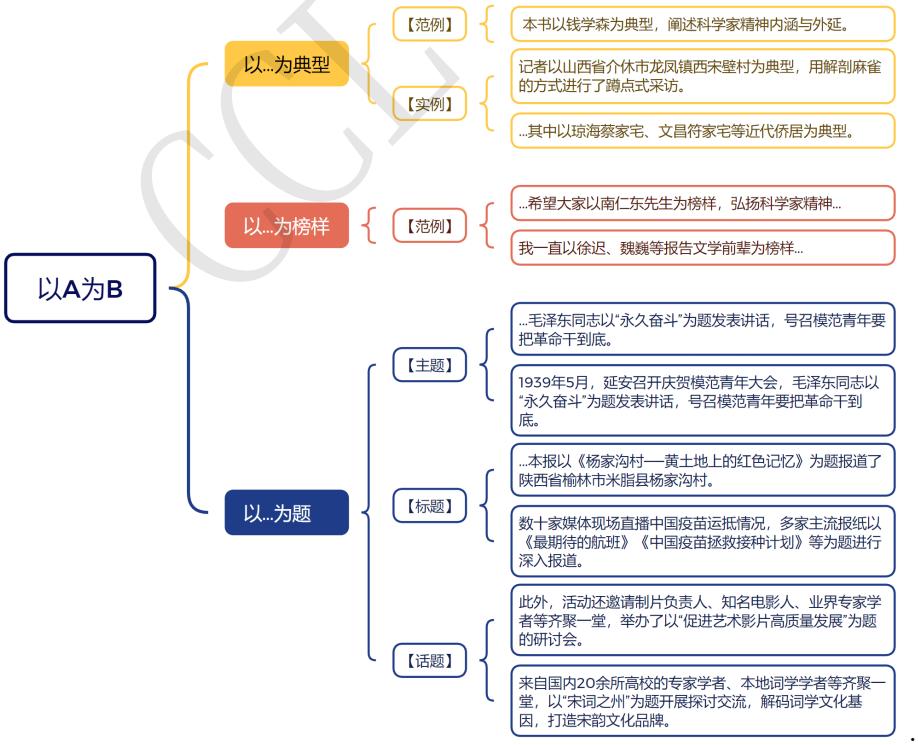



图 5: 构式类型与框架对应关系


(c) 2024 中国中文信息学会计算语言学专业委员会

## 4 以构式为目标词的框架识别

框架语义解析(Frame Semantic Parsing, FSP) 最早由Gildea and Jurafsky (2002)基于英文框架网数据FrameNet提出，是一种细粒度语义分析任务，能表示特定词（目标词）在句子中所能激活的语义场景（框架）以及该场景对应的语义角色。因此，FSP被广泛地应用于机器阅读理解(Guo et al., 2020b; Guo et al., 2020a; 王智强et al., 2016)、文章摘要抽取(Guan et al., 2021a; Guan et al., 2021b)、关系抽取(Zhao et al., 2020)及文本生成(谭红叶et al., 2018) 等多种自然语言处理任务。而框架识别（Frame Identification 以下简称FI）是进行FSP的前提(Su et al., 2021b)，该任务是给定可激活框架的目标词，根据上下文语境，从多个所属框架中选取最符合该目标词语境的语义框架的任务。

该任务的形式化定义如下：给定包含目标词的句子$S$ ，记为$S = (w_1, w_2, ..., w_n)$，其中$w_i$为组成句子的第$i$个词，其中$1 \leq i \leq n$。待识别目标词记为$w_t$，$w_t \in S$。要求通过上下文的语义场景从给定的框架库$F = \{f_1, f_2, ...f_n\}$中选择出合适的框架$f_t$，记为式(1)：

$$f_t = \underset{f_i \in F, w_t \in S}{\operatorname{argmax}} P(f_i | S, w_t) \tag{1}$$

本文中所有框架识别实验均使用accuracy作为评价指标，其中$N'$为所有样本总数，$y_i$为每一个样本的原始标签，$\hat{f}(x_i)$为模型预测结果。具体定义如公式(2)所示：

$$Acc = \frac{1}{N'} \sum_{i=1}^{N'} I\left(y_i = \hat{f}(x_i)\right) \tag{2}$$

### 4.1 模型介绍

**EMARoberta**(Huang et al., 2023) 一种基于多优化策略的框架识别方法，采用多种优化策略以解决模型不稳定的问题，提高模型鲁棒性。采取的优化策略主要有以下三种：(1)指数滑动平均(Exponential Moving Average,EMA)(Klinker, 2011)：给予近期数据更高权重，对模型参数做平均，使得模型参数的更新与一段时间内的历史取值有关，可以提高模型在测试集上的鲁棒性。(2)Warm-up策略：训练开始前先使用一个较小的学习率进行一定的迭代次数，使得模型逐渐适应数据集的特征，让模型的权重更新更加平稳，减少训练时的震荡和不稳定性。(3)FGM对抗训练(Miyato et al., 2021)：对embedding层在梯度方向添加扰动，引入噪声，既提高了模型的泛化能力，又提高了模型的鲁棒性。

**RoPEBert**(Li et al., 2023b) 一种基于旋转位置编码的框架识别方法(Su et al., 2021a)，针对在框架识别中会丢失目标词与整体句子之间的位置信息关系的问题，使用了一种基于旋转式位置编码的方法来计算实体之间注意力信息，然后进行分类和抽取，该方法使用旋转矩阵对绝对位置进行编码，同时将显式的相对位置依赖性纳入自注意公式中。

### 4.2 数据集划分

本研究将之前构建的标注语料数据集按照7：1：2的比例划分为训练集，验证集，测试集。具体情况如表3所示。

| 数据集 | 框架数 | 标注句子数 | 平均长度 |
|---|---|---|---|
| **训练集** | 82 | 16242 | 92.7 |
| **验证集** | 68 | 2367 | 83.8 |
| **测试集** | 80 | 4675 | 84.1 |

表 3: 数据集划分

本文在划分数据集时遵循以下三条规则：(1)训练集中的框架应该在验证集和测试集中出现。(2)相同的框架，但是“B”不相同，一些“B”在测试集中出现，但是在训练集中不出现。(3)相同的“B”但是框架不相同，则有些框架在训练集中出现，有些框架在测试集中出现。


(c) 2024 中国中文信息学会计算语言学专业委员会

采用以上三条规则对本文构建的构式-框架数据集进行划分，测试模型在进行框架识别时是否具备一定的泛化能力和排歧能力，在面对复杂数据集时能否正确的识别出目标框架。

### 4.3 实验结果及分析

本文使用CCL2023-Eval汉语框架语义解析评测任务所提供的基线模型(Li et al., 2023a)，EMARoberta模型及RoPEBert模型作参考，三种模型均采用原论文给定的参数设置。

在本文构建数据集的测试集上，最终实验结果如表4所示。其中Acc是模型所预测的所有测试集数据整体的准确率，AccR2是所有根据划分规则二划分后的数据准确率，AccR3是所有根据划分规则三划分后的数据准确率。

| Models | Acc(%) | AccR2(%) | AccR3(%) |
|---|---|---|---|
| **Bert** | 13.63 | 8.48 | 6.57 |
| **EMARoberta** | 38.26 | **38.29** | 12.39 |
| **RoPEBert** | **42.81** | 17.49 | **32.62** |

表 4: 实验结果

实验结果表明，在目标词为单一词汇的框架识别任务中，表现较好的模型在直接迁移到以构式为目标词的框架识别任务中时表现欠佳。本文选取了部分三种模型都存在的错误进行案例分析，如表5所示。其中，依据划分规则(2)，在划分时将所有“以A为代表”的被标注为【代表】框架的标注实例放入了训练集中，但是没有放入“以A为标志”的被标注为【代表】框架的标注实例。这一规则旨在考察模型在进行框架识别时是否具备泛化能力。实验结果表明，三种模型的泛化能力较弱，导致了模型在寻找“以A为标志”的数据所匹配的框架时，出现了较大的错误率。没有将“以A为标志”的被标注为【代表】框架的数据正确的识别为【代表】框架。而根据划分规则(3)，将“以A为典型”的标注为【实例】框架的标注实例放入了训练集中，但是没有放入“以A为典型”的标注为【范例】框架的数据。这一规则旨在考察模型在进行框架识别时是否具备排歧的能力。实验结果表明，三种模型的排歧能力也较为薄弱，使模型将绝大部分“以A为典型”的数据都笼统的识别为了【实例】框架，没有根据上下文语境对“以A为典型”构式和【范例】框架、【实例】框架进行深入理解，导致了错误的产生。这也与传统的框架识别模型在进行框架识别时，过分依赖目标词所蕴含的语义信息有关。在以构式为目标词的框架识别任务中，构式所蕴含的信息不像单一词汇那样丰富、明显，需要模型根据大量的数据进行学习、挖掘，而这也是现有的框架识别模型所欠缺的。

| 构式类型 | 标注实例 | 模型识别框架 | 正确框架 |
|---|---|---|---|
| **代表** | 近年来，以网络购物、网络娱乐、在线医疗等为代表的数字消费迅猛发展。 | 代表 | 代表 |
| **标志** | 我们相信，以北京冬残奥会的成功举办为标志... | 目标 | 代表 |
| **典型** | 记者以山西省介休市龙凤镇西宋壁村为典型，用解剖麻雀的方式进行了蹲点式采访。 | 实例 | 实例 |
| **典型** | 本书以钱学森为典型人物，阐述科学家精神内涵与外延。 | 实例 | 范例 |

表 5: 部分框架识别结果

### 4.4 面向构式目标词的框架识别方法

传统的利用单一目标词框架识别任务中，目标词为句子中的单一词汇，含有一定的词义信息帮助模型进行识别，但是在以构式为目标词的框架识别任务中，目标词为一种结构，如本文的“以A为B”构式，目标词为“以”，“为”这两个字，本身蕴含的信息较少，不能辅助模型进行框架识别。为了解决这一问题，研究尝试将目标词除“以”，“为”这两个字外，再加入“以A为B”构


(c) 2024 中国中文信息学会计算语言学专业委员会

式中的核心信息“B”。因为“以A为B”构式中的“B”往往与构式的含义有着密切的关系。于是，目标词从仅有“以”和“为”两个字转变为“以”，“为”和“B”，通过这种方法来增强模型对构式的理解，提高模型的排歧能力。

针对解决模型泛化能力不足的问题，本文尝试以“B”为重要提示信息，进行数据增强，向训练集中增加了一定数量“B”的近义词信息，即使用近义词替换“B”中的词汇，不改变其他信息。增强方式如图6所示。

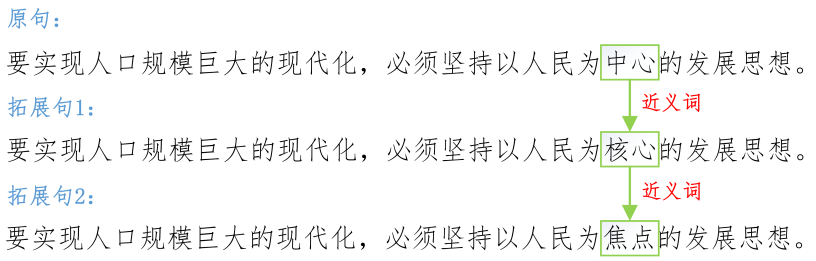



图 6: 数据增强方式

经实验证明，在经过数据增强（Data Augmentation）及目标词转化（Target Word Conversion）后，三种模型的性能均有显著的提高，实验结果如表6所示。这证明了我们提出的两种方法的有效性。

| Models | Acc(%) | AccR2(%) | AccR3(%) |
|---|---|---|---|
| **Bert** | 78.53 | 83.69 | 38.46 |
| **EMARoberta** | **88.19** | **88.17** | 80.76 |
| **RoPEBert** | 82.46 | 80.69 | **88.46** |

表 6: 实验结果

### 4.5 消融实验

为了证明数据增强及目标词转化的有效性，我们进行了消融实验，将数据增强及目标词转化的方法从模型中剔除，最终的实验结果如表7所示。可以得出：(1)将数据增强模块去掉后，模型准确率在Bert模型上下降了27.81%；在EMARoberta模型上下降了12.7%；在RoPEBert模型上下降了4.82%。说明了数据增强方法在以构式为目标词的框架识别任务中的有效性。(2)将目标词转化模块去掉后，模型准确率在Bert模型上下降了45.91%；在EMARoberta模型上下降了28.66%；在RoPEBert模型上下降了17.34%。说明了目标词转化方法能够有效增强模型对句中构式的理解，也证明了目标词转化模块在以构式为目标词的框架识别任务中的有效性。

| Models | -DA(%) | -TWC(%) |
|---|---|---|
| **Bert** | 50.72 | 32.62 |
| **EMARoberta** | 75.49 | 59.53 |
| **RoPEBert** | 77.64 | 62.15 |

表 7: 消融结果

### 4.6 在大模型上的实验

在本文所构建的数据集的基础上，选取了1000条数据进一步在GPT3.5(Ouyang et al., 2022)和Gemini(Team et al., 2023)上进行了测试。实验主要针对Zero-shot和Few-shot两种不同场景设定，利用如图7所示的方法构建了提示模版。在Zero-shot场景下，提示模板中不提供任何带有答案的信息，仅说明任务需求，评估大语言模型自身是否具有框架语义的相关知识，并分析其能否利用相关知识解决框架识别问题。在Few-shot场景下，提示模板中提供了三条带有


(c) 2024 中国中文信息学会计算语言学专业委员会

正确答案的示例样本，使得大语言模型能更好的理解任务需求，从而评估大语言模型能否有效利用少量的示例信息来提升框架识别能力。




图 7: 提示模版构建

实验结果如表8所示。由此可见，面对以构式为目标词的框架识别任务，大语言模型在Few-shot场景下的框架识别能力要显著优于Zero-shot场景，这表明大语言模型具备的上下文学习能力在框架语义分析任务中能够发挥一定作用，能够根据提示样例更好的理解任务需求。然而，无论是在Zero-shot还是Few-shot场景下，大语言模型在以构式为目标词的框架识别任务中的表现与传统模型相比仍存在一定的差异。

| Models | zero-shot(%) | few-shot(%) |
|---|---|---|
| **Gemini** | 23.5 | 37.5 |
| **GPT-3.5** | 12.8 | 21.5 |

表 8: 大模型框架识别结果

## 5 结论

本文以2020年《人民日报》语料库为基础，选取了23849条具有“以A为B”构式的语句作为待标注对象，依据汉语框架标注规范进行了人工标注。最终建立了具有141个框架，超过206万字规模的构式-框架数据集。同时利用了多种框架识别模型及大语言模型进行测试，既验证了数据集的质量，也为框架识别模型的改进提供了一定的参考。验证了大语言模型在以构式为目标词的框架识别任务中距传统模型还有一定差距。文章进一步丰富了基于语言认知的汉语框架语义知识库构建理论，为框架语义知识库扩充和复杂语义场景分析打下一定基础。

然而，本文构建的构式-框架数据集仍然存在一定的局限性，之后将会从以下几点继续改进现有的资源：(1)目前数据集中所覆盖的构式单一，拓展构式种类，增加数据集的多样性；(2)数据集中有较多框架包含的标注实例较少，使用《人民日报》语料库以外的其他语料库，如高考阅读理解语料库等,增加框架下的标注实例数量。

## 参考文献


Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Charles J Fillmore, Paul Kay, and Mary Catherine O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: The case of let alone. *Language*, pages 501–538.

Charles J Fillmore. 1976. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences*, 280(1):20–32.

Charles J Fillmore. 1988. The mechanisms of" construction grammar". In *Annual Meeting of the Berkeley Linguistics Society*, pages 35–55.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational linguistics*, 28(3):245–288.

Yong Guan, Shaoru Guo, Ru Li, Xiaoli Li, and Hongye Tan. 2021a. Frame semantic-enhanced sentence modeling for sentence-level extractive text summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4045–4052.

Yong Guan, Shaoru Guo, Ru Li, Xiaoli Li, and Hu Zhang. 2021b. Integrating semantic scenario and word relations for abstractive sentence summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2522–2529.

Shaoru Guo, Yong Guan, Ru Li, Xiaoli Li, and Hongye Tan. 2020a. Incorporating syntax and frame semantics in neural network for machine reading comprehension. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2635–2641.

Shaoru Guo, Ru Li, Hongye Tan, Xiaoli Li, Yong Guan, Hongyan Zhao, and Yueping Zhang. 2020b. A frame-based sentence representation for machine reading comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 891–896.

Shutan Huang, Qiuyan Shao, and Wei Li. 2023. CCL23-eval 任务3系统报告:基于多任务pipeline策略的汉语框架语义解析(system report for CCL23-eval task 3: Chinese frame semantic parsing based on multi task pipeline strategy). In Maosong Sun, Bing Qin, Xipeng Qiu, Jing Jiang, and Xianpei Han, editors, *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 3: Evaluations)*, pages 105–112, Harbin, China, August. Chinese Information Processing Society of China.

Frank Klinker. 2011. Exponential moving average versus moving exponential average. *Mathematische Semesterberichte*, 58:97–107.

Juncai Li, Zhichao Yan, Xuefeng Su, Boxiang Ma, Peiyuan Yang1, and Ru Li. 2023a. CCL23-eval 任务3总结报告:汉语框架语义解析评测(overview of CCL23-eval task 1:Chinese FrameNet semantic parsing). In Maosong Sun, Bing Qin, Xipeng Qiu, Jing Jiang, and Xianpei Han, editors, *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 3: Evaluations)*, pages 113–123, Harbin, China, August. Chinese Information Processing Society of China.

Zuoheng Li, Xuanzhi Guo, Dengjian Qiao, and Fan Wu. 2023b. CCL23-eval 任务3系统报告:基于旋转式位置编码的实体分类在汉语框架语义解析中的应用(system report for CCL23-eval task 3: Application of entity classification model based on rotary position embedding in chiness frame semantic parsing). In Maosong Sun, Bing Qin, Xipeng Qiu, Jing Jiang, and Xianpei Han, editors, *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 3: Evaluations)*, pages 94–104, Harbin, China, August. Chinese Information Processing Society of China.

Ru Li, Yunxiao Zhao, Zhiqiang Wang, Xuefeng Su, Shaoru Guo, Yong Guan, Xiaoqi Han, and Hongyan Zhao. 2024. A comprehensive overview of cfn from a commonsense perspective. *Machine Intelligence Research*, pages 1–18.

Takeru Miyato, Andrew M. Dai, and Ian Goodfellow. 2021. Adversarial training methods for semi-supervised text classification.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback, 2022. *URL https://arxiv. org/abs/2203.02155*, 13:1.



(c) 2024 中国中文信息学会计算语言学专业委员会

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. 2021a. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*.

Xuefeng Su, Ru Li, Xiaoli Li, Jeff Z. Pan, Hu Zhang, Qinghua Chai, and Xiaoqi Han. 2021b. A knowledge-guided framework for frame identification. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5230–5240, Online, August. Association for Computational Linguistics.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Xiaohui Wang, Ru Li, Zhiqiang Wang, Qinghua Chai, and Xiaoqi Han. 2020. 基于self-attention的句法感知汉语框架语义角色标注(syntax-aware Chinese frame semantic role labeling based on self-attention). In *Proceedings of the 19th Chinese National Conference on Computational Linguistics*, pages 616–623, Haikou, China, October. Chinese Information Processing Society of China.

Liping You and Kaiying Liu. 2005. Building chinese framenet database. In *Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE '05. Proceedings of 2005 IEEE International Conference on*.

Hongyan Zhao, Ru Li, Xiaoli Li, and Hongye Tan. 2020. Cfsre: Context-aware based on frame-semantics for distantly supervised relation extraction. *Knowledge-Based Systems*, 210:106480.

屠寒非, 李茹, 王智强, and 周铁峰. 2016. 一种基于主动学习的框架元素标注. 中文信息学报, 30(4):44–55.

文旭and 司卫国. 2021. 中国构式语法研究20年：回顾与展望. 解放军外国语学院学报, 44(05):43–51+160–161.

李济洪, 王瑞波, 王蔚林, and 李国臣. 2010. 汉语框架语义角色的自动标注. 软件学报, 21(4):597–611.

王智强, 李茹, 梁吉业, 张旭华, 武娟, and 苏娜. 2016. 基于汉语篇章框架语义分析的阅读理解问答研究. 计算机学报, 39:795–807.

詹卫东and 王佳骏. 2022. 面向计算的构式研究：现状、问题与展望. 语言学研究, pages 39–51.

谭世勋. 1985. 试论"以a为b"结构的发展. 华南师范大学学报（社会科学版）, pages 90–96.

谭红叶, 闫真, 李茹, and 敬毅民. 2018. 迈向创造性语言生成:汉语幽默自动生成的探索. 中国科学:信息科学, 48:1497–1509.

魏在江and 赵帮华. 2024. 隐喻式否定构式的语义生成机制研究. 外国语文, 40:21–32.

CCL 2024

# 上古汉语分词和词性标注语料库的构建

**柯永红**
北京师范大学民俗典籍文字研究中心, 中国文字整理与规范研究中心　北京市新街口外大街19号100875
yh5555@126.com

## 摘要

针对国内尚无开放的大规模上古汉语分词及词性标注语料库可用的问题，提出以人工为主+ 机器辅助的标注模式，构建一个包括46部文献的上古汉语分词及词性标记语料库。描述了语料选择、文本分词、词性标注和质量控制等建库过程，分析了该语料库词长、词频、词用等分布，评估了标注质量。已经完成标注的语料库包括323余万字、217万余词。与EvaHan2022基测集和盲测集的分词、词性标注一致度分别为93.70%、89.49%和92.83%、89.86%。该语料库可用于古汉语研究、辞书编撰、语言教学、人工智能等多个领域。

**关键词：** 上古汉语 ； 分词 ； 词性标注 ； 语料库

# Construction of Ancient Chinese Word Segmentation and Part-Of-Speech Corpus

**Yonghong Ke**
Research Center for Folklore, Classics and Chinese Characters,
Research Center for Collation and Standardization of Chinese Characters
No.19, Xinjiekouwai St, Haidian District, Beijing, 100875, P.R.China
yh5555@126.com

## Abstract

In order to solve the problem that there is no open and large-scaleannotated ancient Chinese word segmentation and part-of-speech corpus available in China, a manual and machine assisted tagging model was proposed to construct an Ancient Chinese word segmentation and part-of-speech corpus including 46 documents. The construction process of corpus selection, text segmentation, part-of-speech tagging and quality control are described. The distribution of word length, word frequency, word usage and tagging quality of the corpus are analyzed. The corpus that has been annotated includes more than 3.23 million characters and 2.17 million words. The agreement of word segmentation and part-of-speech tagging with EvaHan2022 base set and blind test set is 93.70%, 89.49% and 92.83%, 89.86%, respectively. The modified corpus can be used in many fields such as ancient Chinese research, dictionary compilation, language teaching, artificial intelligence and so on.

**Keywords:** Ancient Chinese , word segmentation , part-of-speech , corpus

## 1 引言

©2024 中国计算语言学大会
本文为国家社科基金课题“上古汉语词标记语料库及应用系统构建研究”（项目编号：20BYY127）的阶段性成果。

国内尚无开放使用的大规模古籍分词和词性标记语料库，给古籍研究和保护带来极大不便，也限制了人工智能技术在古籍文献上的深层应用。比如，在古籍全文检索语料库输入词条“天然”、“引领”，会检出错误的句子：

例1. 有**天然**后有地，有地然后有别。《鬻子·汤政汤治天下理第七》
例2. 如有不嗜杀人者，则天下之民皆**引领**而望之矣。《孟子·梁惠王上》

例1中“天然”不是一个词。例2中“引领”表示“伸长脖子”，是词组而不是词，而这样的错误需要具备一定的古文阅读素养才能发现，极易对使用者产生误导。

本文的目标是建设一个覆盖上古时期主要文献的分词和词性标记语料库。古籍所对应的古代汉语的面貌与现代汉语有很大不同，其内部各阶段也有较大的分别 (王力, 1980)，本文讨论的是上古汉语分词和词标注语料库的建设。

## 2 研究现状

### 2.1 古汉语分词和词性标记语料库建设现状

我国机读语料库的建设始于1979年。在各界的共同努力下，古籍资源库和语料库的建设取得了丰硕的成果，极大方便了古汉语研究和应用。从整体上看，我国的古汉语语料库建设有三个方面的特点：（1）目前建成的古汉语语料库多是没有分词和词性标记的电子文本语料库，如中华古籍资源库(中国国家图书馆)、中华经典资源库(国家语言文字工作委员会)、中国基本古籍库(北京爱如生数字化技术研究中心)、瀚堂典藏古籍数据库(北京时代瀚堂科技有限公司)、书同文古籍数据库(北京书同文数字化技术有限公司)、国学宝典(北京国学时代文化传播股份有限公司)、籍合网(古联（北京）数字传媒科技有限公司)、国学大师(国学大师)、识典古籍(北京大学-字节跳动数字人文开放实验室) 北京大学CCL语料库(北京大学中国语言学研究中心)、鼎秀古籍全文检索平台(北京翰海博雅科技有限公司)、中国哲学书电子化计划(STURGEON D.)等，这些语料库解决的是古籍文字逐字索引问题。（2）古汉分词和词性标记语料库数量极少，如南京师范大学先秦汉语标注语料库(陈小荷等, 2013)、南京农业大学典籍平行语料库(黄水清, 王东波, 2021)等，尚未完全开放访问。（3）台湾“中研院”古代汉语标记语料库（以下简称ASC）和英国谢菲尔德大学历时汉语语料库（SCC）是世界范围内具备一定规模的、开放的分词和词性标记语料库。ASC的文献数量较多、标注质量高且提供了较为丰富的词检索功能。但是，ASC采用的术语和标签集与大陆存在差异，且有访问限制。SCC的所收语料较少、标签通用性较低且存在一些明显的断句、分词或标注错误。

整体而言，国内尚无完全开放的大规模古汉语分词和词性标注语料库。

### 2.2 古汉语自动分词和词性标注技术研究现状

古汉语自动分词主要有基于词典和统计的方法、基于序列标注的方法和基于深度神经网络的方法。基于深度神经网络的分词和词性标注方法相对于传统方法有明显的优势，俞敬松等 (2020)、Tang Xuemei等 (2022)在部分语料上分词的F值分别为95.32%、98.46%。第一届古代汉语分词和词性标注国际评测中，盲测集上封闭测试分词和词性标注的F值分别达到93.64%和87.77% (李斌, 袁义国, 芦靖雅等, 2023)。总体而言，预训练模型、外部知识引入以及多模型融合方法大大提升了古汉语自动分词的效果。

但是，当下古汉语的自动分词仍然不能满足严谨的科研场景的需求。汉语的大部分复音词是从短语演变而来的，在古汉语阶段，有的复音词还处于由短语向词的转变过程之中，这些复音词身上还带着短语的特征，这给词和短语的区分带来很大困难。北京大学数字人文研究中心“吾与点”是目前最新、最先进的古汉语分词系统之一。以“吾与点”2024年1月的版本为例，在其中输入三组例句，可以得到如下分词结果：

例1. 佑/问/长/：/有/**妻子**/乎/？/对/曰：/有/妻/未/有/子/也/。《后汉书·吴佑传》
例2. **妻子**/好合/，/如/鼓/瑟琴/。《诗经·小雅·常棣》
例3. 是故/君子/有/终身/之/忧/，/无/**一/朝**/之/患/也/。《孟子·离娄下》
例4. 强/而后/可/，/ **一/朝**/而/获/十/禽/。《孟子·滕文公下》
例5. 残贼/之/人/谓/之/**一/夫**/。《孟子·梁惠王下》
例6. 耕者/之/所/获/，/**一/夫**/百亩/。《孟子·万章下》


(c) 2024 中国中文信息学会计算语言学专业委员会

上述分词结果中例4、例6、例8存在错误的分词。例4中的“妻子”表示“配偶和子女”，是词组。例6中“一朝”表“一时”义，是词。例8中的“一夫”表示“众叛亲离之人、独夫”，是词。

我们今天所能见到的古汉语语料是封闭的、有限的、不完整的，一些结构在短语向词转变的过程中存在词和短语两种形态，或者作为词出现的频次很低（表现为未登录词或低频词），导致可用来机器学习的语料不足，这也是限制机器自动分词准确性的主要原因，而这样的问题在可预见的一段时间内仍将存在。古汉语中存在许多兼类、词类活用以及词的临时用法，给词性的标注带来了很多困难。而且，自动分词的错误会叠加词性标注之上。

目前，上古汉语缺乏相应的词性标记规范。《信息处理用现代汉语词类标记规范》(标准号GB/T 20532—2006) 及其修订工作(杨丽姣等, 2021)为现代汉语汉语语料库建设提供了重要依据。化振红(2021)的《建立中古汉语语料库分词规范的若干问题》讨论了中古汉语词的切分及分词规范问题，为上古汉语分词及词性标注提供了重要的参考。

综上所述，在现有技术条件下，古汉语分词和词性标注仍需要大量的专业人工标注和检验工作，这也是建设大规模、高质量古汉语分词和词性标记语料库的主要困难。

## 3 上古汉语分词和词性标记语料库的建设

我们的语料库总体遵循了一套系统化、规范化的建库流程，总体思路可以用下图表示：

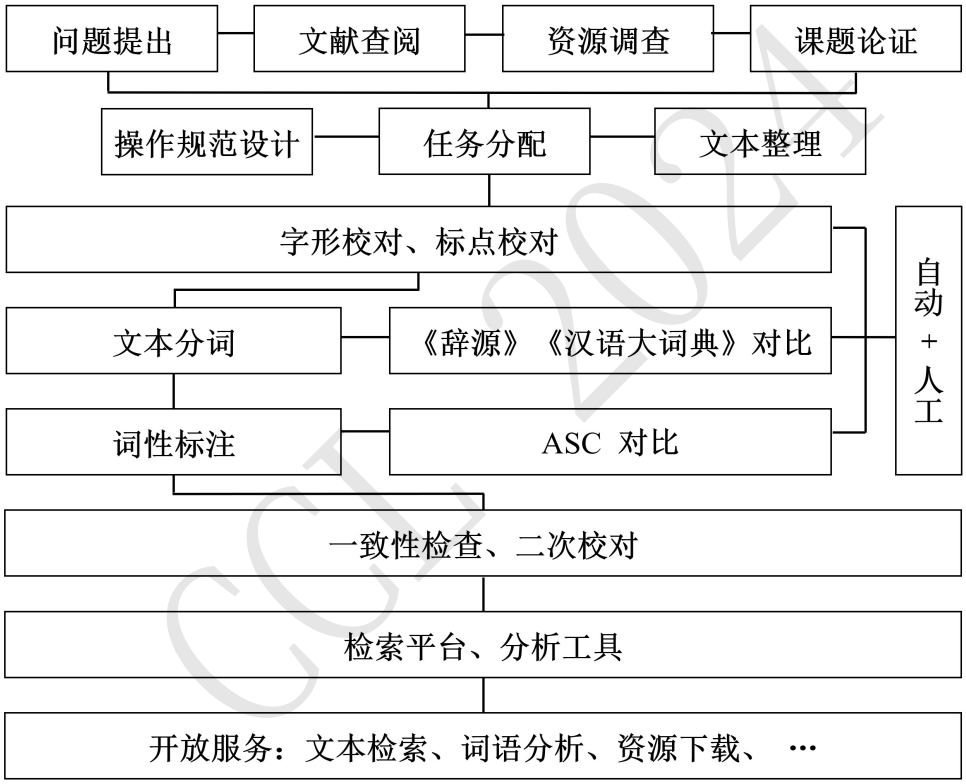



Figure 1: 建库过程

首先，我们通过广泛的文献查阅和资源调查完成课题论证。在确定了研究内容和建库目标后，设计了操作规范，进行任务分配和文本整理，为后续工作做好铺垫。在语料标注阶段，进行了字形和标点校对、文本分词、词性标注等一系列步骤，同时结合自动校对与人工检查的方式进行了比对，以确保数据的准确性和完整性。之后，通过一致性检查和二次校进一步提升语料标注质量。在完成语料标注的基础上，通过检索平台和分析工具，为用户提供包括文本检索、词语分析、资源下载等在内的开放服务。

如前文所述，在现有技术条件下，上古汉语分词和词性标注仍需要大量的专业人工标注和检验工作。为建成高质量的大规模古汉语分词和词性标注语料库，本文采用了自动工具为辅+人工为主的标注模式，主要基于两个方面的考虑：（1）在自动标注阶段，使用机器对专有名词、数词、量词、连词、代词等出现较多且自动标注准确率高的词进行预标注，再由人工


(c) 2024 中国中文信息学会计算语言学专业委员会

对剩余部分进行标注并对机器标注部分进行校对，可以减轻大规模语料标注中人工标注的工作量。（2）在人工标注阶段，充分借鉴已有研究成果，使用《辞源》《汉语大词典》以及ASC等权威辞书和高质量语料库协助人工标注和校对，进一步提升标注准确率。限于篇幅，我们重点介绍语料选取、文本分词、词性标注和质量控制几个方面的工作。

### 3.1 语料选取

上古汉语主要包括先秦和两汉时期。根据李零(2008)的统计，先秦两汉典籍现存115种，其中先秦60种、秦1种、汉54种。我们以文献均衡性和重要性为主要依据，选取了上古时期的46种文献。如表1所示：

| 文献类别 | | 先秦 | 两汉 |
|---|---|---|---|
| 传世文献 | 六艺类（10种） | 《左传》《公羊传》《谷梁传》《诗经》《尚书》《仪礼》《礼记》《大戴礼记》《周礼》《周易》 | |
| | 史书类（3种） | 《国语》《战国策》 | 《史记》 |
| | 子书类（23种） | 《论语》《孝经》《孟子》《荀子》《墨子》《老子》《庄子》《冠子》《慎子》《商君书》《韩非子》《邓析子》《吕氏春秋》《管子》《晏子》《孔子家语》《孔丛子》 | 《新语》《新书》《新序》《淮南子》《春秋繁露》《说苑》 |
| | 兵书类（5种） | 《司马法》《六韬》《孙子》《吴子》《尉缭子》 | |
| | 方技类（2种） | 《素问》《灵枢》 | |
| | 经传小学类（1种） | | 《韩诗外传》 |
| 出土文献 | （2种） | 马王堆帛书、睡虎地秦简 | |

Table 1: 语料库所包含的文献

上述文献已经包含了上古时期的主要文献，能够满足日常使用需求。随着研究工作的进展，我们还在增加新的文献，最终的目标是覆盖目前已有的全部文献。

### 3.2 文本分词

将句子切分成词串是建设分词和词性标记语料库核心工作。但是，古汉语特别是上古汉语的分词与现代汉语相比，情况要复杂得多。在古汉语中，有大量的复音词处于词和词组的中间状态。同样一个结构在现代汉语是词，在古汉语有时是词有时是词组。比如“大国”“大雨”和“大海”“大我”，“大国”和“大雨”意义为“大的国”“大的雨”，与之相应的还有“小国”“小雨”，应当是词组。“大”是“海”自身的属性和特点，没有“小海”之说，“大海”应当是词。“大我”尽管有对应的“小我”，但其中的“大”和“小”不和“我”直接发生关系，应当看作是词(王宁, 2008)。马真(1981)指出“确定先秦时期的合成词，根据的只是有限的书面材料，不能像区分现代汉语的词和词组那样可以采用‘替换法’‘扩展法’‘插入法’等，更不能简单地使用是否经常连用作为定词的标准”。具体操作而言，我们遵循几个主要的分词原则：

(1) 意义方面，根据语素组合后是否表达一个整体的意义来判定是否为复音词。构成词的两


(c) 2024 中国中文信息学会计算语言学专业委员会

个或多个语素组合后表达一个整体的意义，且不同于参构语素按照结构关系推导出的意义(卜师霞, 2018)。语素组合一般有产生新义、意义泛化或具体化、意义偏指或构成偏义等几种情况。

例1. **左右**曰：“待先生，如此其忠且敬也。”《孟子·离娄下》
例2. 予死于**道路**乎？《论语·子罕》
例3. 古者**大事**，必乘其产。《左传·僖公十五年》
例4. 王曰：“舍之！吾不忍其**觳觫**，若无罪而就死地。”《孟子·梁惠王上》
例5. 则易直子谅之心**油然**生矣。《礼记·乐记》

例1中“左右”作为一个整体构成了新的意义，例2中“道路”构成了一个更加概括的意义，例3中“大事”的意义具体化，特指战争。例4中“觳觫”表达一个不可拆分的、整体的意义，例5中“然”作为词缀不表义。因而，上述五个例子加粗部分都应当划分为词。

(2)语法方面，根据语素组合结构的稳固性判定是否为复音词。构成复音词的两个或多个语素组合应结合稳定，不能拆开或者随意扩展，否则是词组。或者组合结构的多个成分中，有一个一般不单用。

例6. 退食自公，**委蛇委蛇**。《诗经·羔羊》
例7. 昔我往矣，杨柳**依依**。《诗经·采薇》
例8. 子曰：“父母之年，不可不知也。”《论语·里仁》
例9. 鲍叔牙为人，刚愎而上悍。刚则犯民以暴，愎则不得民心。《论语·里仁》
例10. 他人有心，予**忖度**之。《孟子·梁惠王上》

例6和例7中“委蛇”“依依”不能拆开，否则没有意义，因而是词。例8中“父母”可以扩展为“父与母”，是词组。例9中“刚”和“愎”在后面的句子中以单用，因而前句的“刚愎”是词组。例10中“忖”表“揣度”义时，一般都不单用，因而“忖度”应作为一个分词单位。

(3)根据复音词与其构词语素的词性变化判定是否为复音词。毛远明(2000)指出词义的变化和词性的转变是同步的，词性转变了，应该是复音词。

例11. 凡祀，**启蛰**而郊。《左传·桓公五年》
例12. 孔张，君之昆孙，子孔之后也，**执政**之嗣也。《左传·昭公十六年》

例11中“启”和“蛰”都是动词，“启蛰”变为一个名词作状语，词性发生了转变，因而可以判定为复音词。例12中“执政”本为动宾结构，组合后变为名词性成份，可以判定为复音词。

(4) 根据上古汉语单音词占据绝对优势的特点，对于尚未完全固定的语素组合，坚持从严的标准，能分则分。

例13. 王何必曰利？亦有**仁义**而已矣。《孟子·梁惠王上》
例14. **仁义**忠信，乐善不倦。《孟子·告子上》

尽管“仁”和“义”作为组合出现的频次很高，但“仁”和“义”均可单独使用，且组合后意义也未发生改变，因而组合尚未完全固定，应切分为两个单位。

(5)同一语素组合有时是词，有时是短语，需要根据具体语境判定是否属于一个分词单位。

例15. 苟行王政，**四海**之内皆举首而望之。《孟子·滕文公下》
例16. 禹之治水，水之道也。是故禹以**四海**为壑。《孟子·告子下》

上面两个句子中，前一个“四海”表“天下”之义，是一个词，后面的“四海”表“四方之海”之义，是词组。

(6) 借助权威工具书和已有电子资源检验分词结果。我以《辞源》和《汉语大词典》作为复音词的检验参照、以ASC校对词性标注，开发了辅助工具在标注过程中为标注者提供提示性信息。上述工具书和电子资源是提示性的，不作为标准。

### 3.3 词性标注

王力(1990)指出“在划分词类的时候，不但要重视结构方面（形态方面），而且要重视意义方面。应该把结构和意义看成一个有机的整体。”划分词类时，我们遵循是功能和意义结合的标准。和现代汉语相比，古汉语的词类划分更加复杂，具体而言，需要重点关注兼类、词类活用和词的临时用法：

（1）古汉语中词的兼类非常普遍，需要在语境中依据意义和功能进行判断。以“使”为例：

第二十三届中国计算语言学大会论文集，第819页–第829页，太原，中国，2024年7月25日至28日。
卷1：主会论文
(c) 2024 中国中文信息学会计算语言学专业委员会

例17. **使**勿使能殖，则善者信矣。《左传·隐公六年》
例18. 公怒，绝宋**使**。《左传·隐公九年》
例19. **使**能，国之利也。《左传·文公六年》

上面三个“使”分别应当分别标注为动词、名词和副词。

（2）古汉语中的词类活用也远多于现代汉语，常有名词作动词、形容词作动词、数词作动词等词类活用现象出现。比如“东”本来是名词，在：

例20. 秦师遂**东**。《左传·僖公三十年》

中“东”活用作为动词使用。在词性标注过程中，需要根据具体语境判断，这要求标注者有很好的古代文献阅读素养。

（3）此外，古汉语中还有一些词的临时用法，应当参照语境义划分词类。如：

例21. 天地车**轮**。《吕氏春秋·大乐》

高诱注：“轮，转。”据殷国光(2008)的统计，“轮”在春秋战国文献中出现的34例用法均为名词，义为“车轮”。殷国光认为此例中“轮”作“转”是一种临时意义，应以“常功能为标准”，将其划分为名词。我们认为殷国光关于词义的分析是正确的，但是在归类时，我们以“语境义为标准”，将其归入动词，这么做一方面符合该词的实际用法，同时也可以在语料库中保留更多的语言现象，有利于语言研究。

上古汉语缺乏相应的词性标记规范，在进行语料标注时，研究者通常会建立一套自用的标注集，但总体而言，关于词的大类划分的区别不大。杨伯峻、何乐士(2001)在《古汉语语法及其发展》中提出了一套全面、细致的词类划分方案，我们以两位先生的研究为主要参照，并考虑与《信息处理用现代汉语词类标记规范》(标准号GB/T 20532—2006, 标准号GB/T 20532—2006)尽量兼容，将词类划分十三个大类。词类和对应的标记如下表所示：

| 序号 | 标记 | 词类 | 词例 |
|---|---|---|---|
| 1 | a | 形容词 | 大 |
| 2 | c | 连词 | 而 |
| 3 | d | 副词 | 渐 |
| 4 | e | 叹词 | 呜呼 |
| 5 | j | 兼词 | 诸 |
| 6 | m | 数词 | 一 |
| 7 | n | 名词 | 东 |
| 8 | o | 拟声词 | 关关 |
| 9 | p | 介词 | 以 |
| 10 | q | 量词 | 杯 |
| 11 | r | 代词 | 吾 |
| 12 | u | 助词 | 者 |
| 13 | v | 动词 | 学 |

Table 2: 词类和标记

### 3.4 质量控制

由于自动分词还不能很好地区分上古汉语的词与非词，我们采用了人工为主+自动工具为辅的标注模式。根据李斌等(2012)对25部先秦文献的统计，名词、动词和人名是数量最多的词类。专有名词、数词、量词等数量多且自动标注准确率高，我们使用自动标注工具对这几类词进行预标注，然后由教师、研究生、本科生组成的标注小组标注剩余部分并对预标注内容进行检查。小组标注完成后再开展二次人工校对，校对工程中发现的难点、疑点问题由项目团队专家团队通过会议评审解决。过程可由下图表示：

标注人员的前期培训、分词与词性操作手册、古籍注疏参考资料都是必要的质量控制过程和材料。我们也开发了一些辅助工具来减少人工标注的错误，如标注标签自动校对和自动提示等。下图展示了利用外部资源自动校对的功能：


(c) 2024 中国中文信息学会计算语言学专业委员会

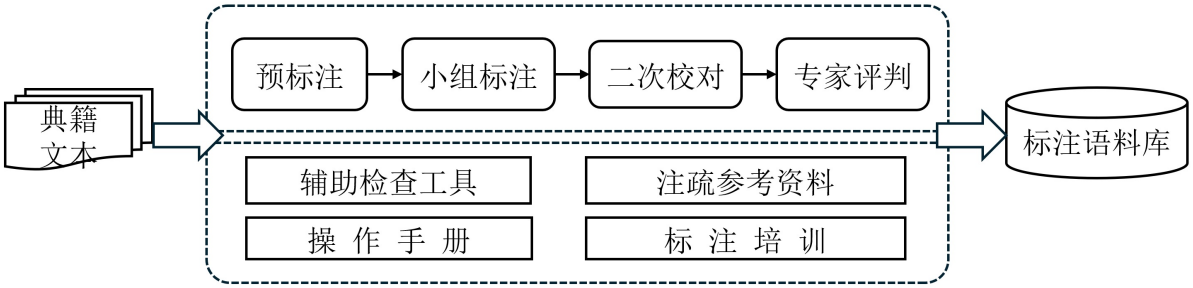



Figure 2: 建库过程

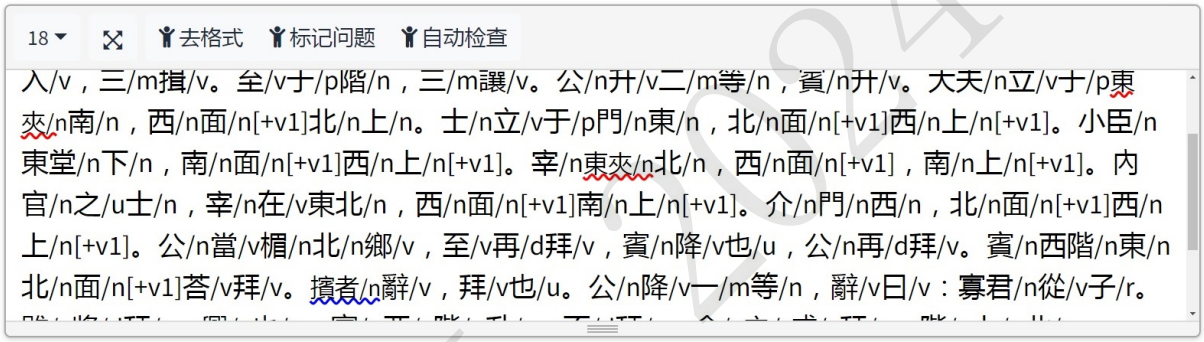



Figure 3: 辅助标注工具

上图中红色波浪线表示与ASC的分词不一样，蓝色波浪线表示《汉语大词典》中未收录此复音词，提示标注者注意检查。随着完成语料的增加，我们也将逐步完善辅助标注工具，让人工标注+ 智能辅助的方式取得更高的质量和效率。

## 4 上古汉语分词和词性标记语料库内容、质量及应用

### 4.1 语料库的内容

自2020年课题启动以来，我们已经完成46部典籍的标注工作，其中17部已经完成二次校对。根据统计，语料库的词长如下表所示：

| 词长 | 词数 | 百分比% | 词形数 | 使用率 |
|---|---|---|---|---|
| 1 | 1907750 | 87.69 | 6925 | 275.49 |
| 2 | 237473 | 10.92 | 37506 | 6.33 |
| 3 | 22758 | 1.05 | 6265 | 3.63 |
| 4 | 5670 | 0.26 | 2419 | 2.34 |
| 5 | 1473 | 0.07 | 872 | 1.69 |
| 大于5 | 372 | 0.02 | 331 | 1.12 |

Table 3: 语料库的词长分布


(c) 2024 中国中文信息学会计算语言学专业委员会

王宁(1996)指出，汉语词汇的积累经历了原生阶段、派生阶段到以双音合成词为主的合成阶段。从上表的统计来看，呈现了先秦及两汉阶段汉语词汇的特点：从词的数量来说，先秦文言文中的单音词占据绝对优势，明显多于复音词；从词的使用频率来说，单音词的使用频率明显高于复音词。我们发现，三字词四字词多是专名（如周幽王、楚隐王、古公甫、强弩都尉、二十八宿、骇鸡之犀）、数字（如七十一、十分之二）以及一些特殊结构（如隐隐兮、焉、、暖暖姝姝），词长为5及以上的基本都是数字（如三千七十一、二百五十八、四百六十余）。

我们将语料库中的词频按照词性排序，可以得到下表：

| 序号 | 词性 | 数量 |
|---|---|---|
| 1 | 名词 | 733885 |
| 2 | 动词 | 627803 |
| 3 | 助词 | 186071 |
| 4 | 副词 | 160679 |
| 5 | 代词 | 145181 |
| 6 | 连词 | 106758 |
| 7 | 形容词 | 94693 |
| 8 | 介词 | 75649 |
| 9 | 数词 | 37919 |
| 10 | 量词 | 4778 |
| 11 | 兼词 | 1657 |
| 12 | 叹词 | 423 |
| 13 | 象声词 | 372 |

Table 4: 语料库的词性分布

可以看出，名词、动词、助词数量最多。上古汉语代词中副词、代词的占比高于现代汉语，副词在上古阶段仍处于虚化的过程中，代词一部分属于实词、一部分属于虚词。

### 4.2 语料库的标注质量

我们将该语料库（下文简称ACP）与第一届古代汉语分词和词性标注国际评测（EvaHan2022）发布的数据集（下文简称EVA）(李斌, 袁义国, 芦靖雅等, 2023)进行了对比。EVA语料主要包括《左传》《史记》和《资治通鉴》。《资治通鉴》成书于北宋，我们计算时将此部分排除，保留《左传》和《史记》。将ACP、ASC的标注数据与EVA进行了句子级别的对齐后，以EVA为基准，进行F1值的计算，结果如下表所示：

由于ASC、ACP及EVA的词性标记存在较大差异，我们只计算了ASC的分词指标，没有计算其词性标注指标。以ASC为基准，则ACP分词F1为95.59%，可见分词方面，二者的一致性很高。我们重点分析了ACP与EVA的标注差异。标注差异主要有三种情况。首先，二者对分词粒度的把握不同，如：

例1. ACP：及/c **宋人/n** 盟/v 于/p 宿/n
例2. EVA：及/c **宋/n 人/n** 盟/v 于/p 宿/ns


(c) 2024 中国中文信息学会计算语言学专业委员会

| | 数据集 | 分词F1 | 词性标注F1 | 备注 |
|---|---|---|---|---|
| EvaHan2022最佳指标 | 基测集 | 96.16% | 92.05% | 封闭测试 |
| | 盲测集 | 93.64% | 87.77% | 封闭测试 |
| ASC 指标 | 基测集 | 92.50% | / | / |
| | 盲测集 | 92.90% | / | / |
| ACP 指标 | 基测集 | 93.70% | 89.49% | / |
| | 盲测集 | 92.83% | 89.86% | / |

Table 5: ACP、ASC与EVA 标注一致性对比

例3. ACP：**郑/n 公子忽/n** 为/v 质/n 于/p 周/n
例4. EVA：**郑公子忽/nr** 为/v 质/n 于/p 周/n

上例中，可以看到ACP和EVA对名词不同的处理方式。由于名词出现频次最高，这样的情况占据了二者差异的主要部分。在EVA中，我们也可以看到对“郑公子忽”这一单位不同的处理方式：

例5. EVA：**郑/ns 公子忽/nr** 在/v 王所/n

其次，ACP和EVA对一些结构有不同的判定，如：

例6. ACP：庄公/n **寤/a 生/v** ，/w 惊/v 姜氏/n ，/w 故/c 名/v 曰/v 寤生/n 。/w
例7. EVA：庄公/nr **寤生/v** ，/w 惊/sv 姜氏/nr ，/w 故/c 名/n 曰/v 寤生/nr 。/w 例句中出现了两次“寤生”。“庄公寤生”中“寤生犹言逆生，现代谓之足先出”(杨伯峻, 2001)，“寤生”应当标注为短语，ACP的标注更为恰当。第二个“寤生”作为人名出现，ACP和EVA都标注为名词。又如：

例8. ACP：去/v 顺/n 效/v 逆/n ，/w **所/u 以/p** 速/v 祸/n 也/u 。/w
例9. EVA：去/v 顺/n 效/v 逆/n ，/w **所以/c** 速/sv 祸/n 也/y 。/w

王力(1989)指出“所以”完全变为连词大约是从晋代开始，杨伯峻、何乐士(2001)认为“‘所以’作为一个连词，则大约是汉以后的事。”“所以”在上古时期是两个词，现代汉语的“所以”是由介宾结构逐步虚化而来的。因而，ACP将上例中的“所以”标注为两个词。

最后，ACP和EVA对词类及标记的划分并不一致，ACP的词类标记有13个，EVA的词类标记有21个，尽管我们已经将词标记进行了对齐，但这也在一定程度上影响了词性标注F1的计算。

上古汉语的词的划分历来是难点，实际标注过程中也有不同的操作标准，因而导致各家的标注结果会有所不同。对于难以判定的语素组合，ACP比EVA的分词粒度更小，ACP采取能分则分的策略。总体而言，我们认为，ACP的分词粒度符合上古汉语的实际情况，词性标注准确，能较好地满足科研和应用的需要。

### 4.3 语料库的应用

上古汉语分词和词性标记语料库不仅可以提供更加准确、灵活的词检索和统计，研究者更可以对定量模式进行定性的功能解释，对于汉语词汇的意义、结构和用法研究包括词汇面貌的整体描写、词义分析、词用调查、类型调查、形态特征分析等有重要的作用。

近年来“大数据”“深度学习”等前沿技术对于现代的汉语研究和应用起到了很大的推动，研究人员越来越渴望借助计算机对古籍开展全面、深层的研究。上古汉语分词和词性标记语料库可以作为一个基础的资源用作机器训练和学习，作为前沿技术服务于古汉语研究和应用的支撑性资源。

我们正在开发语料库检索平台和统计分析工具，未来将开放给语言研究、语言教学、辞书编撰、机器翻译等领域使用。


(c) 2024 中国中文信息学会计算语言学专业委员会

## 5 结语

语料库加工的本质是语言知识的重新整理、发现、形式化、规范化等工作，是最基本、最重要的基础研究之一。论文介绍了上古汉语词性标注语料库的构建过程的思路、难点和过程，并从词汇词频、词性和词长和分析了语料库的分布。论文的研究有助于填补学界缺少开放的大规模、深标注古汉语语料库的空白。论文的特色之处在于采用了人工标注为主+ 自动化工具辅助的模式，对词和词组的处理较为准确。下一步，将从三个方面继续开展工作：增加语料库的内容，最终实现对上古文献的全覆盖；继续对已经标注的语料开展人工和自动校对，提高标注的一致性和准确性；分批发布语料，共享给学术界和工业界使用。语料库加工的本质是语言知识的重新整理、发现、形式化、规范化等工作，是最基本、最重要的基础研究之一。论文介绍了上古汉语词性标注语料库的构建过程的思路、难点和过程，并从词汇词频、词性和词长和分析了语料库的分布。论文的研究有助于填补学界缺少开放的大规模、深标注古汉语语料库的空白。论文的特色之处在于采用了人工标注为主+ 自动化工具辅助的模式，对词和词组的处理较为准确。下一步，将从三个方面继续开展工作：增加语料库的内容，最终实现对上古文献的全覆盖；继续对已经标注的语料开展人工和自动校对，提高标注的一致性和准确性；分批发布语料，共享给学术界和工业界使用。

## 参考文献

GB/T 20532-2006, 信息处理用现代汉语词类标记规范[S]. 北京: 中国标准出版社, 2007.

STURGEON D. *Chinese Text Project*[EB/OL]. [2023-04-01]. https://ctext.org/

TANG X M, SU Q. 2022. *That Slepen Al the Nyght with Open Ye! Cross-era Sequence Segmentation with Switch-memory*, volume 1. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7830–7840, Dublin, Ireland. Association for Computational Linguistics.

北京爱如生数字化技术研究中心. 中国基本古籍库[EB/OL]. [2023-04-01]. http://dh.ersjk.com/

北京大学中国语言学研究中心. *CCL*语料库检索系统[EB/OL]. [2023-04-01]. http://ccl.pku.edu.cn:8080/ccl_corpus/index.jsp

北京大学-字节跳动数字人文开放实验室. 识典古籍[EB/OL]. [2023-04-01]. https://www.shidianguji.com/

北京国学时代文化传播股份有限公司. 国学网[EB/OL]. [2023-04-01]. http://www.gxbd.com/

北京翰海博雅科技有限公司. 鼎秀古籍全文检索平台[EB/OL]. [2023-04-01]. http://103.242.200.9/ancientbook/portal/index/index.do

北京时代瀚堂科技有限公司. 瀚堂典藏古籍数据库[EB/OL]. [2023-04-01]. https://www.hytung.cn/

北京书同文数字化技术有限公司. 书同文古籍数据库[EB/OL]. [2023-04-01]. http://guji.unihan.com.cn/

卜师霞. 源于先秦的现代汉语复音词初探[M]. 北京: 中华书局, 2018.

陈小荷, 冯敏萱, 徐润华等. 先秦文献信息处理[M]. 北京: 世界图书出版公司, 2013.

古联（北京）数字传媒科技有限公司. 籍合网[EB/OL]. [2023-04-01]. https://www.ancientbooks.cn/

国家语言文字工作委员会. 中华经典资源库[EB/OL]. [2023-04-01]. https://zhjdzyk.eduyun.cn/

国学大师. 国学大师[EB/OL]. [2023-04-01]. http://www.guoxuedashi.net/

化振红. 建立中古汉语语料库分词规范的若干问题[J]. 语言研究集刊, 2021, (06): 151-167.

黄水清, 王东波. 国内语料库研究综述[J]. 信息资源管理学报, 2021, 11(03): 4-17+87. DOI: 10.13365/j.jirm.2021.03.004.

李斌, 冯敏萱, 陈小荷. 基于标注语料库的先秦汉语词汇统计[R]. 武汉: 第十三届汉语词汇语义学研讨会, 2012.


(c) 2024 中国中文信息学会计算语言学专业委员会

李斌, 袁义国, 芦靖雅等 . 第一届古代汉语分词和词性标注国际评测[J]. 中文信息学报, 2023, 37(03): 46-53+64.

李零. 简帛古书与学术源流（修订本）[M]. 北京: 三联书店, 2008.

马真. 先秦复音词初探[J]. 北京大学学报(哲学社会科学版), 1981, (5): 77.

毛远明. 左传词汇研究[M]. 重庆: 西南师范大学出版社, 2000.

王力. 汉语史稿[M]. 北京: 中华书局, 1980.

王力. 汉语语法史[M]. 北京: 商务印书馆, 1989.

王力. 王力文集*(*第十六卷*)*[M]. 山东教育出版社, 1990: 315.

王宁. 当代理论训诂学与汉语双音合成词构词研究[A]. 沈阳, 冯胜利主编. 当代语言学理论和汉语研究[C]. 北京: 商务印书馆, 2008.

王宁. 训诂学原理[M]. 北京: 中国国际广播出版社, 1996.

杨伯俊, 何乐士. 古汉语语法及其发展（修订本）[M]. 北京: 语文出版社, 2001.

杨伯峻. 春秋左传注（修订本）[M]. 北京: 中华书局, 2016.

杨丽姣, 肖航, 刘智颖. 《信息处理用现代汉语词类标记规范》修订研究[J]. 语言文字应用,2021,(03):111-120.

殷国光. 《吕氏春秋》词类研究[M]. 北京: 商务印书馆, 2008.

俞敬松, 魏一, 张永伟等 . 基于非参数贝叶斯模型和深度学习的古文分词研究[J]. 中文信息学报, 2020, 34(06): 1-8.

中国国家图书馆. 中华古籍资源库[EB/OL]. [2023-07-01]. http://read.nlc.cn/thematDataSearch/toGujiIndex

CCL 2024

# 图解句式结构体系及其树库构建

彭炜明[1,2,*]，赵敏[3]，宋昱辰[4]，胡佳佳[1]，宋天宝[5]，穗志方[6]，宋继华[1]
[1] 北京师范大学汉字汉语研究与社会应用实验室，中国北京，100875
[2] 宾夕法尼亚大学语言数据联盟 (LDC)，美国费城，19104-2653
[3] 北京汉雅天诚教育科技有限公司，中国北京，100084
[4] 密歇根大学语言学系，美国安娜堡，48109-1220
[5] 北京工商大学计算机学院，中国北京，100048
[6] 北京大学计算语言学教育部重点实验室，中国北京，100871
pengweiming@bnu.edu.cn, zhaom@han-sky.com, songyuch@umich.edu
hjj81@126.com, songtianbao@btbu.edu.cn, szf@pku.edu.cn, songjh@bnu.edu.cn

## 摘要

句式结构是一种基于句本位语法的形式化句法结构，采用自定义的图解形式呈现句子结构。本文提出了涵盖小句结构、词法结构和句间结构三方面的句式结构体系，阐明了其设计理念以及句本位的析句原则，最后概述了基于该体系构建汉语树库的工程进展情况。

**关键词：**句式结构；句本位语法；图解法；汉语树库

# Diagrammatic Sentence Pattern Structure System and Its Treebank Construction

**Weiming Peng[1,2,*], Min Zhao[3], Yuchen Song[4], Jiajia Hu[1], Tianbao Song[5], Zhifang Sui[6], Jihua Song[1]**
[1] Laboratory for Chinese Character and Language Research and Social Application, Beijing Normal University, Beijing, 100875, China
[2]Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA, 19104-2653, USA
[3]Beijing Han-sky Education Technology Co., Ltd, Beijing, 100084, China
[4]Department of Linguistics, University of Michigan, Ann Arbor, MI, 48109-1220, USA
[5]School of Computer Science and Engineering, Beijing Technology and Business University, Beijing 100048, China
[6]Key Laboratory of Computational Linguistics, Ministry of Education, Peking University, Beijing; 100871, China
pengweiming@bnu.edu.cn, zhaom@han-sky.com, songyuch@umich.edu
hjj81@126.com, songtianbao@btbu.edu.cn, szf@pku.edu.cn, songjh@bnu.edu.cn

## Abstract

The Sentence Pattern Structure (SPS) is a formalized syntactic structure based on sentence-based grammar, which presents the structure of sentences in a customized diagram form. This paper proposes an SPS system covering three aspects: clause structure, lexical structure, and inter-sentence structure, elucidating its design philosophy and parsing principles of sentence-based grammar. Finally, an overview of the engineering progress of building a Chinese treebank based on this system is provided.

**Keywords:** Sentence Pattern Structure; Sentence-based Grammar; Diagrammatic Analyzing Method; Chinese Treebank

*通信作者：彭炜明（pengweiming@bnu.edu.cn）
基金项目：国家自然科学基金（62007004）；北京市自然科学基金（423408）；国际中文教育研究课题（23YH35C）。
©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

## 1 引言

句式结构体系是基于黎锦熙《新著国语文法》所创“句本位”语法理论而设计的一种形式化句法结构体系。由于它采用自定义的“图解法”来表现句子结构，故也可称作图解句式结构。图解法的数字化改造工作最早可追溯至（彭炜明，2012），其目的是将曾经流行一时的汉语语法教学工具付诸计算机实现，以推动汉语语法教学和中文信息处理两个领域的信息沟通。其后，我们实现了数个版本的图解方案和可视化工具，构建过规模不等的句式结构树库（也叫句本位语法树库）。今年是第一个以“句式结构”（Sentence Pattern Structure）命名的图解方案公布的 10 周年（彭炜明等，2014），由于不同时期的设计不尽相同，因此有必要做一个阶段性总结，同时发布最新版的句式结构体系及其树库语料。

## 2 相关工作

汉语的树库构建起步于 20 世纪 90 年代，数十年间建成了一批影响较大的中文树库。先是以宾州中文树库 (CTB)（Xue N., 2005）、清华汉语树库 (TCT)（周强, 2004）为代表的短语结构树库，其后是以哈尔滨工业大学中文依存树库 (HIT-CDT)（Che W., 2012）、苏州大学汉语依存树库 (CODT)（郭丽娟, 2019）为代表的依存结构树库。这两种结构各有侧重，前者突出短语功能和结构层次，后者强调中心词和结构关系，它们是目前中文信息处理领域的主流句法结构。北京大学的多视图中文树库 (PMT)（邱立坤, 2015）尝试融合二者优点，实现了两种结构的自动切换。

由于主流树库大多是以计算机自动句法分析为背景，其结构体系是在英语成熟的句法结构体系基础上改造而成的，在面对汉语文本时表现出一些不适应性。近年来，汉语树库研究出现了如下一些新的动向：

- **句法标注向浅层分析转向。**即避免繁琐标注过程，不追求以词为单位的分析粒度，而改以组块为单位，以“呈现句子骨干结构”为目标，例如北京语言大学的汉语组块依存树库（钱青青, 2022）。
- **突破句法约束直接标注语义结构。**应对汉语中大量存在的非投影结构，采用更为自由的图（而非树）来表示句子的语义结构。例如哈尔滨工业大学的语义依存图库（Wang Y., 2016）、南京师范大学的中文抽象语义表示语料库 (CAMR)（李斌, 2017）。
- **分析单位从句子延伸至段落乃至篇章。**针对汉语多流水句和“意合”的语法特点，树库建设开始关注语句衔接和篇章结构问题。例如苏州大学的微观和宏观篇章结构语料库（奚雪峰, 2017; 褚晓敏, 2020）和北京语言大学的小句复合体语料库（宋柔, 2017; 2022）。
- **面向语言教学回归传统教学语法。**为适应面向人的语言教学需求，采用传统教学语法描写句子结构。例如北京师范大学的黎氏语法树库（He J., 2013）、句本位图解树库（Song T., 2016），以及北京语言大学自动构建的句式结构树库（谢晨晖, 2023）。

以上有些是相互关联的，比如，北语组块依存树库进一步发展为基于篇章的句法结构树库（卢露, 2022）、意合图表示（荀恩东, 2023），就体现了前三个方面。而（4）其实就是前三个动向的综合，传统教学语法在突显句子结构骨架，处理句法语义的结构冲突，以及促进段落篇章理解方面有着同样的诉求。

本文所谓的传统教学语法即以“句子成分”分析（而非“直接成分”分析或层次分析）为基础，以归纳句型句式为主要目标的汉语语法体系。《暂拟汉语教学语法系统》（张志公, 1956）及此前的多数汉语语法体系大都可归入这一类，最典型的即黎锦熙“句本位”语法。由于黎氏语法的句子成分划分与今天通行的教学语法存在较大差异，因而后来的句式结构体系设计（Peng W., 2015）便以《暂拟系统》作为主要参照。

## 3 句式结构体系

句式结构分析首先由程序根据截句标点[0]（。｜？｜！）自动切分出“整句”，整句之下再分出若干小句。小句内部的成分结构以二维平面上自定义的图解样式呈现，各小句的图解样式再以不同虚线相连，表示各种句间结构。词语散布在图解样式上，标注词类、义项和词法结构信息。

[0]截句标点的范围也可视文本情况扩充到冒号、分号，以及一部分的省略号、破折号。


(c) 2024 中国中文信息学会计算语言学专业委员会

总之，一个完整的句式结构分成小句结构、词法结构和句间结构三个层面。图解操作在图形界面上进行，由程序完成图解样式和数据格式（XML）间的编、解码转换。

### 3.1 小句结构

句式结构共设主、谓、宾、定、状、补和独立语七类句子成分，虚词不单独充当句子成分而占据特定“虚词位”。如图 1 所示，句子成分沿一条长横线平铺排列：主、谓、宾为主干成分，置于横线上方；定、状、补、独立语为附加成分，附于横线下方。主干维持“主-谓-宾”单谓语核心（简称单谓核）格局的句式为“基础句式”；打破这一格局的为“复杂句式”，其中联合谓语[1]、连动谓语、兼语谓语和合成谓语[2]为双谓核结构，主谓谓语结构则以支架顶起，形成一个新的句式层次。

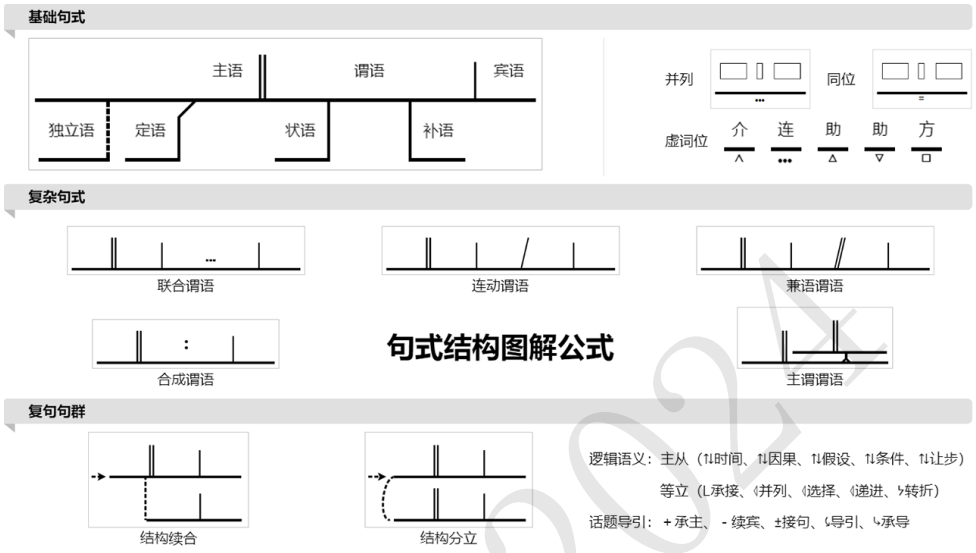



Figure 1: 句式结构图解公式

虚词位置于其所连接或附加的成分旁边，分四类：介词位、连词位、方位词位和助词位，前三类分别标示“∧”、“...”、“□”于横线下方，助词位又分三种：①连接定、状、补语的助词“的/之”、“地”、“得”，置于折线与横线的连接处；②与 NP 或单词组合的助词，标示“△”于横线下方；③与 VP 或小句组合的助词，标示“▽”于横线下方。图解公式中各横线段均可进行横向扩展，但遵循如下约束规则：

- (1) 主、宾语为体词性横线段（简称体词段），谓语为谓词性横线段（简称谓词段）[3]；
- (2) 体词段只能扩展出定语、介词位、方位词位、△ 助词位、并列结构、同位结构；
- (3) 谓词段只能扩展出状语、补语、宾语、主语、▽ 助词位、连词位、双谓核结构；[4]
- (4) 附加成分横线段初始不限定体/谓性质，当其进行结构扩展时，再根据 (2) 和 (3) 确定；
- (5) 超出 (2) 和 (3) 限制的扩展，如主谓谓语、VP 主语等，需在原横线段上用支架顶起以造成新的结构层次；

句式结构有其独特的层次观。若只对主干成分进行规则 (2) 和 (3) 范围内的扩展，只会引起主干横线的延长，并不增加句式层次。增加句式层次的情况如图 2 所示：

[1]联合谓语：两个 VP 按复句逻辑紧缩而成，例如：母亲求神拜佛。他越想越生气。
[2]合成谓语：助动词/系动词+VP，例如：谁能帮助我？他昨天是从香港回来的。
[3]句本位语法不设“名词谓语句”，一般所谓“名词谓语”分析为缺省谓语“是”所带的判断宾语。
[4]许多从谓词段扩展出的成分并不与其直接相连，如主前状语、宾后补语，但应视为从谓词段而出。


(c) 2024 中国中文信息学会计算语言学专业委员会

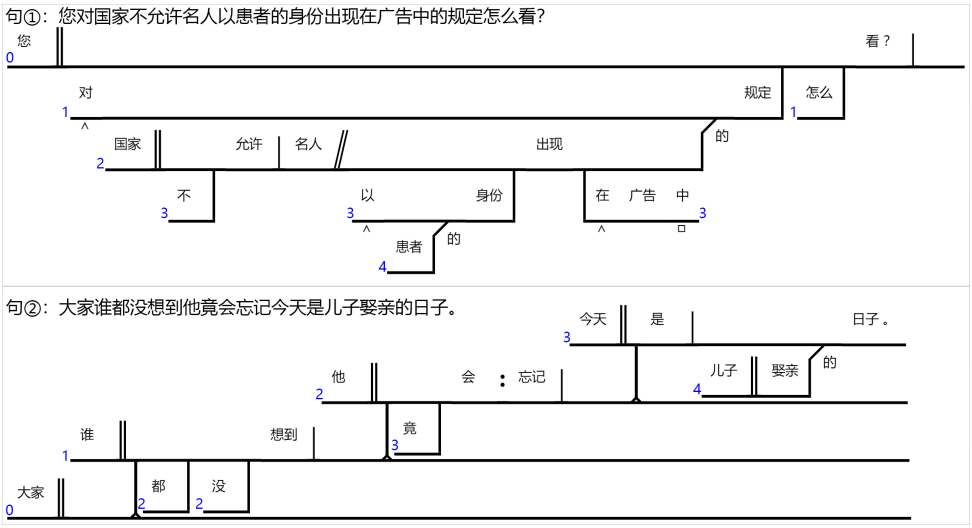



Figure 2: 句式结构的层次深度示例

### 3.2 词法结构

成分分析一般到词一级结束，但由于汉语词和短语边界不清，图解时就以一个基础词库为界，凡词库中收录的称作“词库词”，未收录而又不宜进一步句法切分的单位统称“动态词”。词库词直接标注词类和义项；对动态词，先从内部拆分出词素（不必拆至最小的语素，以能找到词库义项为界），分别标注词素类、义项、词法结构关系（如表 1）以及动态词的整体词类。

| 符号 | 关系 | 示例 |
|---|---|---|
| … | 并列 | 桌…椅；中…小学 |
| ↗ | 定中 | 鸡 ↗ 蛋；小 ↗ 白 ↗ 兔；群众 ↗ 接待 ↗ 站 |
| → | 状中 | 极 → 具；深 → 感；代 → 写；改 → 用 |
| ← | 述补 | 赶 ← 跑；看 ← 清；拿 ← 下；举 ← 起 |
| \| | 动宾 | 驯 \| 兽 □ 师 |
| ‖ | 主谓 | 你 ‖ 争…我 ‖ 夺 |
| • | 重叠 | 看 • 看；研究 • 研究；睡 • 睡觉；看 • 一 • 看；看 • 了 • 看；看 • 不 • 看 |
| - | 其它词法 | 桌-上；两-只；一-大-碗；看-了；同学-们；学习-者；拿-得-起；翩然-而-至 |
| ◇ | 超词法 | 十-元 ◇ 八-角 ◇ 六-分；江西 ↗ 省 ◇ 宁都 ↗ 县 ◇ 固村 ↗ 镇 |

Table 1: 词法结构关系

词法标注示例如图 3，XML 中为动态词整体设一个词类节点，词素类作为其孩子节点，词素类节点上设属性@sen 记录义项，词类节点上设属性@mod 记录如下 3 方面信息构成的词法结构模式：词素类、词素字数（为 1 时省略）和词法结构关系。

### 3.3 句间结构

篇章中的句子并非孤立存在，而是通过指称、结构衔接、逻辑连接等手段组织起来的。汉语语法具有“意合”特点，小句之间的结构衔接表现出显著的“流水句”特性。参考宋柔提出的广义话题结构及其堆栈模型，句式结构在处理小句的结构衔接时也引入缩进对齐的描写机制，在图解样式上分为两类（参见图 1）：

- **结构续合**：后续小句共享前面小句的句首成分序列，图解样式上缩进后续小句，并以虚直线相连，对齐至共享部分的末端，相当于（宋柔, 2013）中的堆栈模型。蒙后共享的情况即（宋柔，2013）中的逆向堆栈模型，也归此类，但因数量较少，本文阐述均以承前共享为准。
- **结构分立**：不满足上述续合关系的其它衔接方式，图解以虚弧线相接两个小句主干横线的最左端。其中，后一小句必无缩进；前一小句若缩进，则以虚线延伸其主干横线至最左端。


(c) 2024 中国中文信息学会计算语言学专业委员会

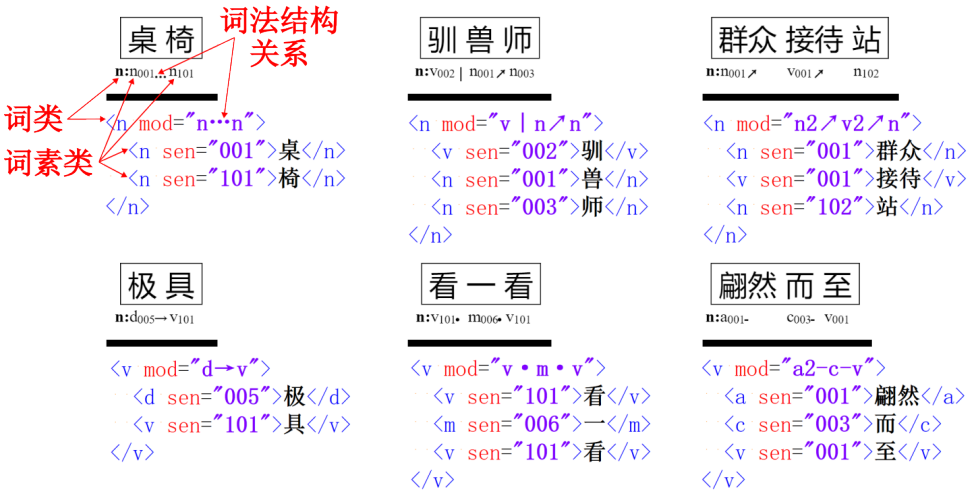



Figure 3: 动态词的词法结构及 XML 格式

例如图 4。XML 中，小句结构和词法结构编码在 xj 节点之下（限于篇幅此处省略，具体参见 (Peng W., 2015)），句间结构通过 xj 的属性@upt（up to 的缩写，意为“向上对齐至”）表明。取值格式为“m:n”，其中 m 表示相连 xj 的 id，n 表示接续至该 xj 的第 n 个词之后，若“结构分立”则取值“s”。结构续合仅限于同一整句内部，而分立小句则可能位于不同整句，此时@upt 中加上表示整句 id 偏移量的 k 值（一般取负值），格式为“k:m:n”（详后图 5）。

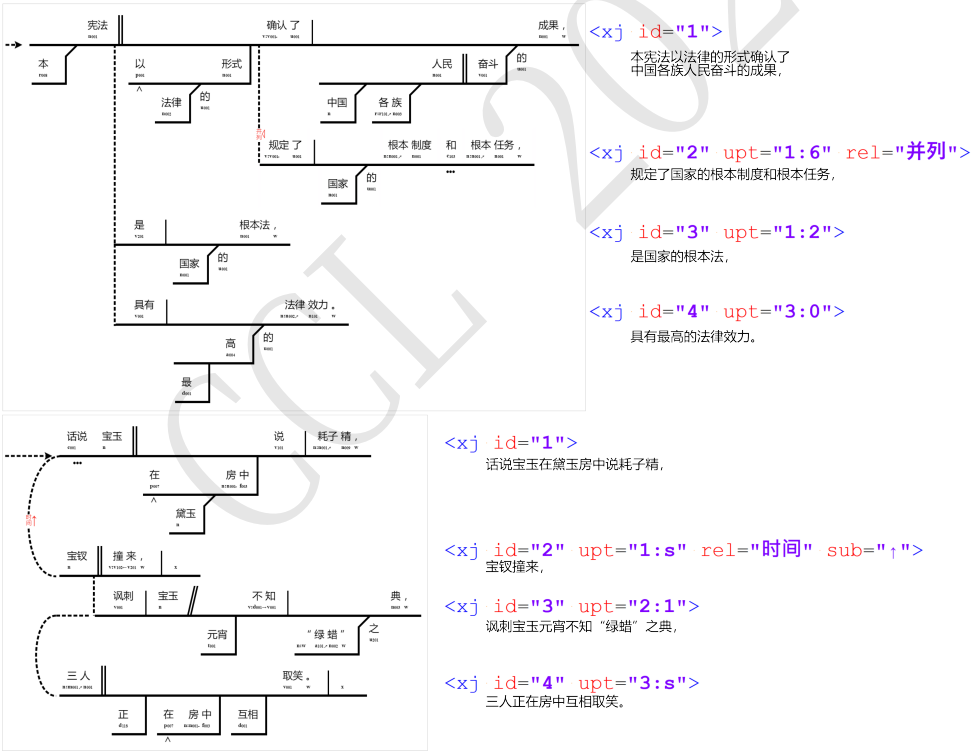



Figure 4: 句间结构的续合和分立

相连小句之间若存在逻辑语义关联，标注于虚线上，在 XML 中记录于 xj/@rel 属性。参照（黎锦熙, 1962）复句分类体系，句式结构中逻辑语义关系共分两大类、十小类：

- 等立：承接、并列、选择、递进、转折；
- 主从：时间、因果、假设、条件、让步。


(c) 2024 中国中文信息学会计算语言学专业委员会

主从复句中，由于从句一方可前可后，故在关系旁加标“↑”或“↓”，以示从句所在位置（对应 XML 属性@sub）。仿此，等立关系旁也增加一个标示符，粗分为“承接 L、并列 ⦃、选择 ⦃、递进 ⦃、转折 ϟ”。由于“承接”可视为小句相连的缺省关系，故若无形式特征（关联词等），可以不显式标注。

小句间的“结构续合”是语法上经济原则的作用，即通过共享，后面小句避免了重复前面小句中出现的句首成分序列。然而，小句之间的共享信息并不限于此，比如（宋柔, 2013）中的“新支句/节栈模型”，即后面小句以前面小句的非开头部分作“话题”。这种情况均以“结构分立”的虚弧线图解，但加标“话题”关系：

- 承主（＋）：后小句话题取自前小句的主语，或从更前小句继承来的话题。
- 续宾（－）：后小句话题取自前小句的宾语，包括状、补语中的介词宾语。
- 接句（±）：前小句整体作为后小句的话题。

此外，在涉及引语（包括直接引语和间接引语）的篇章中，引语后面的小句经常跨越引语部分（有时很长，涉及多个小句）而从前面的导语小句中延续话题。为此，再引入两类“话题导引”关系：

- 导引（↳）：前、后小句分别为导语、引语，当引语涉及多个小句时，只标与导语相邻的小句。
- 承导（↳）：后小句话题取自前小句对应导语小句的话题。

图解样式中，“导语”关系的虚弧设为有向弧，而其它 4 类则加标记于虚弧线旁，对应 XML 编码设在 xj/@tpc 属性中，例如图 5 所示。

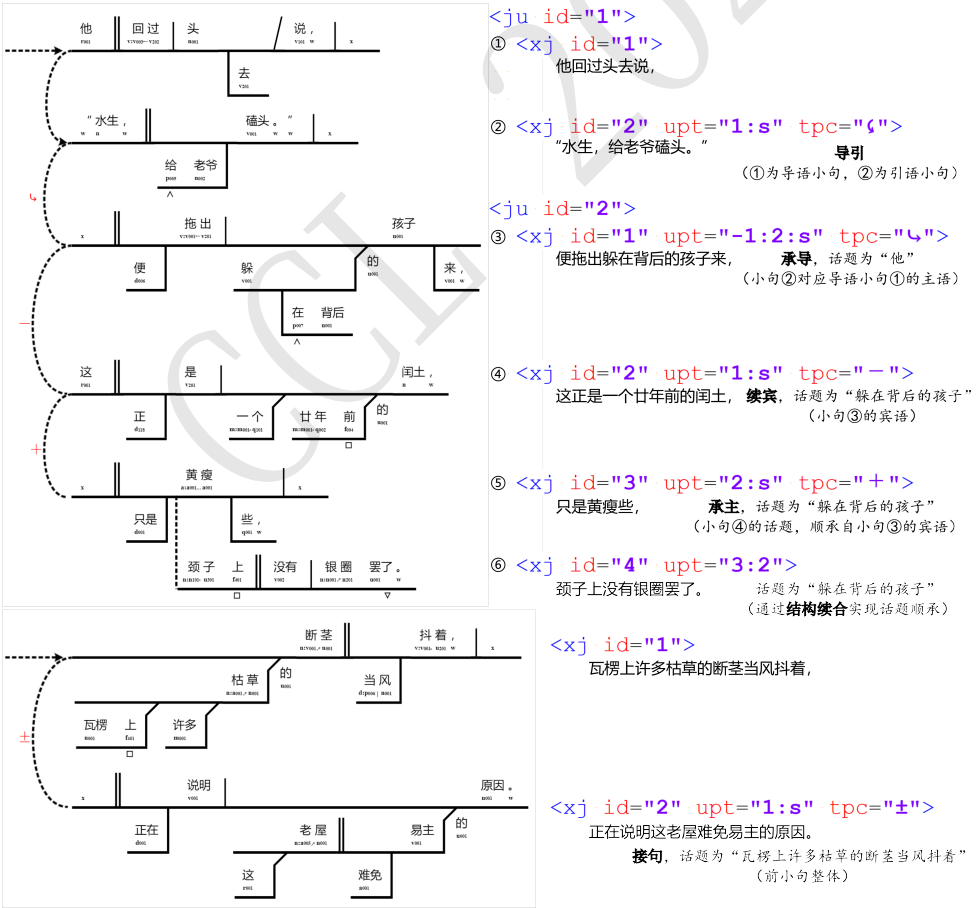



Figure 5: “话题导引”示例

CCL 2024

## 4 设计理念和析句原则

句本位语法是一种教学语法，与强调分析程序严密性的结构主义语法相对。将图解法改造为句式结构体系，始终需要面对的一个问题就是兼顾其教学性质和信息化需求。从结构主义的角度看，句本位语法有许多不利于信息处理的特点，有些恰恰是其教学性质的突出反映，有着心理认知基础和教学实用目的。本节将从坚持教学性和实现信息化这一对矛盾的平衡处理入手，阐述句式结构体系的设计理念，并探讨该体系下句本位的析句原则。

### 4.1 直观呈现，提高空间利用

句式结构采用自定义的图解样式表达句子结构，直观性是其区别于现有形式化句法结构的最显著特征。以短语结构和依存结构为背景，句式结构的直观性来自于：

- 充分利用二维平面的空间性，不同层次深度的词在横、纵两个维度展开排布，通过横、竖、虚、实等线条实现层次对齐；而主流树库局限于“树”的数据结构，也受制于原文词序列的线性排列，空间利用率都不如句式结构。

- 以长横线分隔主干成分和附加成分，实现语法分析“抓主干”的实用目的。语法教学中最基本的“句子成分”在主流树库中隐藏在二元的层级结构或句法关系中，而在句式结构中它们是直接存在的实体，因此，获取句子主干和成分信息在句式结构中更为便捷。

- 图形样式和数据存储分离，由程序实现编、解码，标注人员只面对图形而不接触数据，避免了记忆标记集符号的负担。

从空间利用率角度出发，也就不难理解句式结构的如下设计：

- 缩进对齐手段的使用仅限于严格的结构续合，而没有如广义话题结构那样扩展至所有的流水模型。因为句式结构是完全的结构分析，若将“话题导引”关系也一律采用缩进形式，则会造成大片的空白区域。

- 在“话题导引”中，虚弧相接遵从就近原则。例如图 5 中，本来小句③可以从①中直接“承主”，小句⑤可以从③中直接“续宾”，但如此就会形成跨小句的虚弧线。“承导”关系的引入就是为了避免当引语小句较多时导语和引语后面小句之间的超长虚弧。

### 4.2 廓清层面，突显句式层次

句式结构的句式层次有别于结构主义语法的短语层次。直观看，短语层次倾向于二分，而句式层次则更加扁平化。事实上，句式结构层次观的独特性还首先表现在将不同层面的结构层次从句式层次中分离出来，具体为：

- 引入动态词，将词法结构的内部层次从句法层面剥离。由于动态词的词法组合与句法有着较高的一致性，主流树库中通常不做严格区分，进而如重叠、数量、词缀结构等词法结构也进入句法层面。在二分为主的层次结构中，这并不构成太大问题；但句式结构需要突显句式层次，讲究“抓主干”。如不设词法结构这一层，则介于词和短语之间的大量动态词结构将混入句法层，势必造成层次嵌套过于细碎，不利于突显主干和句式。

- 小句之间的结构续合和话题导引也具有一定的层次性，本质上都可归结为跨小句的广义话题统摄关系。它们在句式结构的描写机制中都没有采用类似短语结构的层次嵌套，就是为了使小句内部的结构分析能够维持在成分组合（包括虚词位）这一级别。

廓清了上述两个不同层面的层次，便能计算句式结构的分析层次。在单一句式层次的前提下，句子成分根据图解公式遵从特定的组合模式（彭炜明，2021），比如，不带补语的单谓核单宾语句式为：

- 连? 状* 定* 主 连? 状* 谓 定* 宾 助?（其中“?、*”为正则元字符）

已知特定成分序列，如“定-定-主-状-状-谓-定-宾”，便可推定其短语结构层次：

- [[ 定-[ 定-主 ]]-[[ 状-[ 状-谓 ]]-[ 定-宾 ]]]

第二十三届中国计算语言学大会论文集，第830页-第840页，太原，中国，2024年7月25日至28日。
卷1：主会论文
(c) 2024 中国中文信息学会计算语言学专业委员会

因此，句式结构不对短语层次做显式的描写，一些短语层次的结构歧义也不进行消歧。例如“他老板和哥哥是同学”的析句结果仅为“(他) 老板 .. 和 . 哥哥 ‖ 是 | 同学”，其中定语辖域存在歧义。这确是句式结构层次的不足，但换个角度看，该句本身确实存有歧义，即所谓“真歧义”，因而不消歧也不失为一种存疑的做法。

此外，话题导引中“接句”和“导引”两类不按“结构续合”分析（如将后小句缩进对齐至前小句的最末端），除了空间利用率的原因外，也有结构层次的考虑：这两类结构中前后小句的成分序列不在同一句式层次。

### 4.3 依句辨品，建立正则规则

“依句辨品”，即根据词语充当句子成分的类型来辨别词类，是句本位语法的词类主张。在句式层次约束下，从句式到成分可以构造有限的正则规则；那么，“依句辨品”就是进一步构造从成分到词这一级规则，进而建立起完整析句规则系统的要求。

句式结构秉承句本位语法的词类观，力求在成分和词类之间维持一个相对稳定的对应关系：主、宾语对应体词，谓语对应谓词。现实中的矛盾来自于词库词类与语料标注词类的不一致，最典型的就是主、宾语中动词的名物化问题。我们主张从词库词到句中词之间有一个转类过程，当单个谓词进入图解中的体词段时，便转为名词。

根据我们的标注经验，词库（以《现汉》为例）的词类信息未能涵盖真实语料中句法功能的情况不在少数。并且对于同一类型的转类，有的词条为两类单独设置义项，有的则仅设一类，不尽统一。因此，拘定标注语料中的词类都来自词库的做法也不现实。

转类看似多一道手续，但收益大于麻烦，具有现实可行性。一来，主、宾语对应体词、谓语对应谓词的情况占多数，少数情况执行转类就可建立起完整的析句规则系统。二来，除名物化外，还有其它诸如“名 ⇄ 量”、“动 ⇄ 介 ⇄ 连”等的转类，均有规律可循，黎氏语法中有一套系统的转类理论（黎锦熙，1959）可以借鉴。

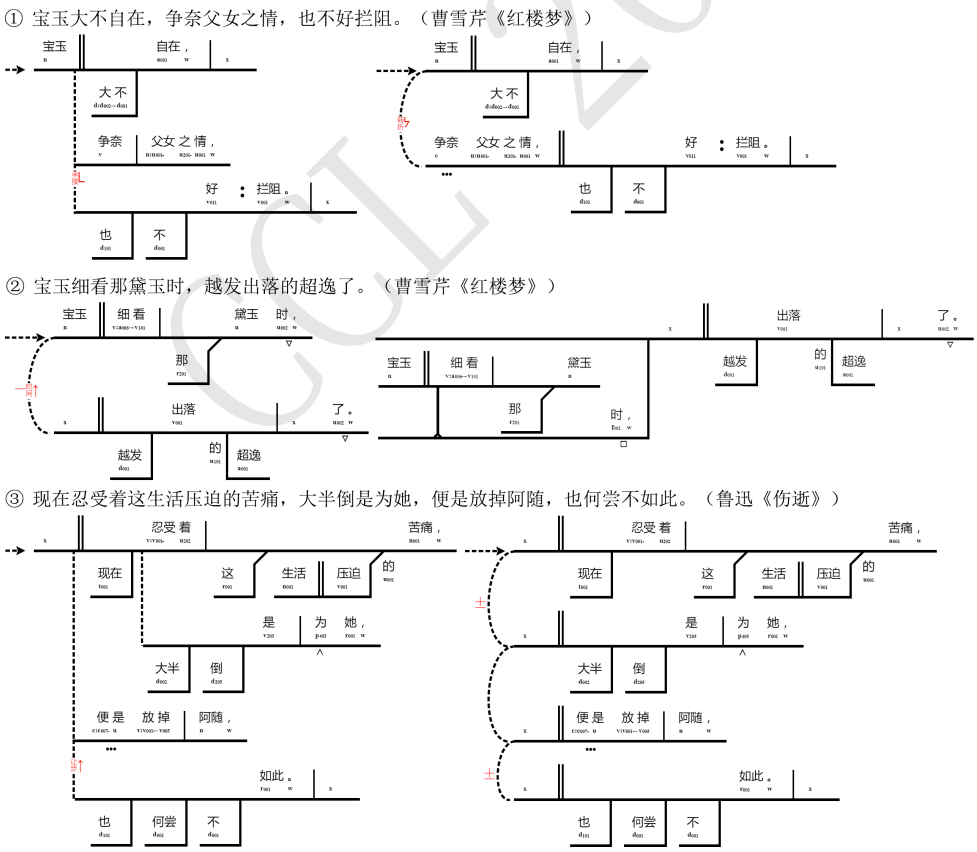



Figure 6:“两可”的结构分析


(c) 2024 中国中文信息学会计算语言学专业委员会

### 4.4 服务章法，活看句法结构

句本位语法图解的最终目标是走到篇章结构分析，也就是说，词法分析、句法分析都只是手段，都是为“章法”服务的。句本位语法著述（黎锦熙,1959;1962）经常提到“活看”，即同一句子作“两可”的结构分析，根据上下文语境有不同取舍。试举几例，如图 6 所示：

例①左图断作三小句，以“续合”相接；右图断作两个“分立”小句，形成“转折”的逻辑关联。例②左图分析为“时间”关系主从复句，右图则分析为带时间状语的单句。句中“争奈”、“时”的词类认定也随句法结构而变。例③左右断句相同，两种分析源于对第二、四小句主语/话题的不同认定，左图以人（省略的“我”）为话题，右图以事（前一小句）为话题，两种分析逻辑上都能自洽。取舍需看更大范围的上下文，着眼于汉语表达的“简明、连贯、得体”，这就是所谓“章法”了。

## 5 树库工程

图解句式结构提出至今，我们已开发出多个版本的语料标注工具（杨天心, 2014; 赵敏, 2014; Guan, 2020）。由于体系结构一直处于升级调整状态中，树库语料的发布滞后于实际的构建进程。截至目前，共标注了 30 余万句规模的现代汉语树库和 2.5 万句规模的古汉语树库。北京师范大学语言与文字资源研究中心、汉字汉语实验室、北京大学计算语言学研究所（ICL）、北京汉雅天诚教育科技有限公司、宾夕法尼亚大学语言数据联盟（LDC）等单位先后参与构建。表**??**为目前树库的语料分布情况。最新一版包含 5000 余句古今篇章语料的树库已提交至 LDC 发布，后续语料也将以篇章或书籍为单位陆续发布。

| 语言 | 类别 | 文本来源 | 规模（句） | 进度 (√: 完成;Δ: 进行中) | | |
|---|---|---|---|---|---|---|
| | | | | 小句 | 词法 | 句间 |
| 白话 | 二语教材 | 《长城汉语》等国际汉语教材 | 140,833 | √ | √ | Δ |
| 白话/文言 | 语文教材 | 高中语文教材 | 12,015 | √ | √ | √ |
| 白话 | 文学作品 | 《毛泽东选集》<br>《鲁迅选集》<br>《红楼梦》<br>《四世同堂》等 | 174,953 | √ | √ | Δ |
| 白话 | 学生作文 | 中学生考场作文 | 5,590 | √ | √ | √ |
| 文言 | 经典古籍 | 《左传》《论语》《孟子》 | 25,388 | √ | √ | Δ |

Table 2: 树库语料分布

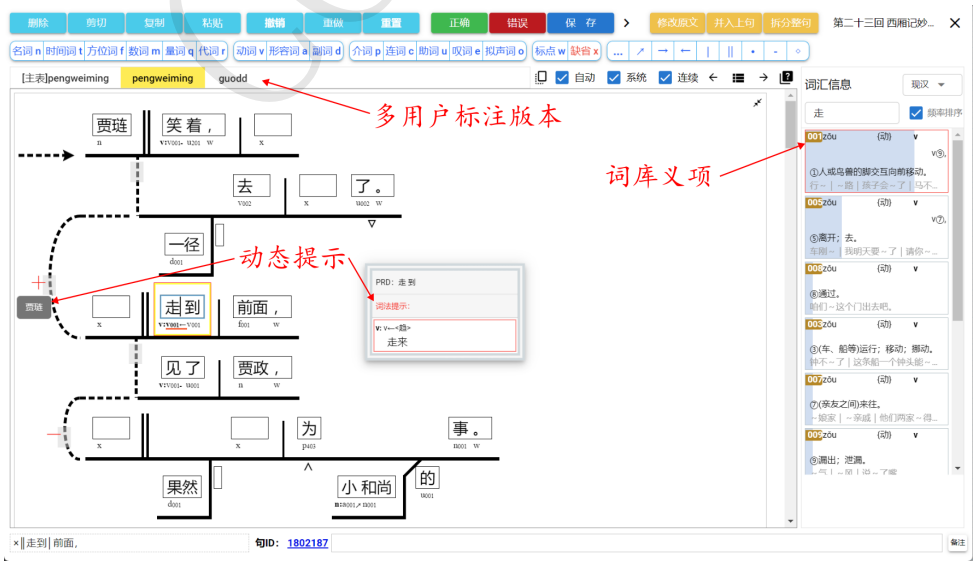



Figure 7: 图解标注系统界面


(c) 2024 中国中文信息学会计算语言学专业委员会

最新版的图解标注系统地址为：http://www.jubenwei.com/，标注界面如图 7 所示。

图解系统背后集成了基础词库、补充词库、词法结构模式库、句式结构规则库、关联词语模式库等一系列知识库，可以动态提示词法、句法结构以及话题导引信息，因而树库标注的效率持续提升。新版树库在增加句间结构信息的情况下，平均标注效率仍能达到 350 句/天的较高水平，略高于（Guan, 2020）的 345 句/天。

由于图解结构相对直观的层次性以及由此而来“抓主干”效果，句式结构的正误判断一目了然。权衡标注成本和质量两方面因素，我们没有选择众包、多人盲标的试验模式来验证一致性，而是采用一标多审的质量控制机制：一位标注员首先标注至副表（个人版本），再经多人共同审核、校改后存入主表。审校人员由本团队内部相对稳定的专职人员充当，分为主审和陪审两种角色。由经验丰富的主审员主持审校，从多位陪审员中选择一人负责操作，其他人同步观看并提出不同意见。陪审员只有取得和主审员超过 80%的整句一致率（即小句、词法、句间完全一致）时才可成为主审员。当遇到“两可”标注意见时，记录后由专家仲裁。疫情期间，我们还组织专家和审校人员对部分经典篇章进行过小规模线上集体标注。如果“两可”标注也计入“一致”的范围，则审校人员在小句、词法、句间这三方面的平均一致率分别可达 92%、95%、84%。

此外，由于我们的语料取材多来自教材和文学作品，并且标注、审校不只是单一枯燥的句法操作，而是综合了词汇、语法、章法的语篇学习过程，因而用户体验要优于一般的树库标注。这也是多人共审和集体标注能够持续开展的原因之一。

## 6 结语

句子结构的图形表示早在上世纪 80 年代的析句方法大讨论中就有人做过对比分析，吕冀平（1984）、廖序东（1986）都指出图解法的优点，“它清晰，醒目，犹如工业上的制图，在讨论句子结构时，双方只要把自己的理解用图解表示出来，就可以省去许多语言。如果要对很长的句子进行分析，那末这种图解法几乎成为**唯一能够使用**的手段，其他各种图示都不免黯然失色。”可惜的是，这种教学上的实用工具在后来很长时间内未能实现信息化，以至错失了在中文信息处理中发挥应有作用的机会。在人工智能和大语言模型盛行的当下，句法分析在计算语言学研究中的关键核心地位受到很大冲击，能否与面向人的语言教学结合起来或是一条值得探索的出路。本文研究正是在这一方向上做出的积极尝试，我们清楚，目前的体系设计中多有迁就传统教学语法，而在信息处理特别是自动分析方面考虑不周的地方，诚恳希望得到有关专家的指正。

## 参考文献

Che W., Li Z., Liu T. 2019. Chinese dependency treebank 1.0 (LDC-2012T05) [DB/OL]. Philadelphia: Linguistic Data Consortium, https://catalog.ldc.upenn.edu/LDC2012T05.

Guan S., Peng W., Song J., et al. 2020. The Construction of Interactive Environment for Sentence Pattern Structure Based Treebank Annotation. Proceedings of 20th Chinese Lexical Semantics Workshop (CLSW2019), Beijing.

Guo D., Zhu S., Peng W., et al. 2016. Construction of the Dynamic Word Structural Mode Knowledge Base for the International Chinese Teaching. Proceedings of 17th Workshop on Chinese Lexical Semantics (CLSW 2016), Singapore.

He J., Peng W., Song J., et al. 2013. Annotation schema for contemporary Chinese based on JinXi Li’s grammar system. Proceedings of 14th Workshop on Chinese Lexical Semantics (CLSW2013), Zhengzhou.

Peng W., Song J., Sui Z., et al. 2015. Formal Schema of Diagrammatic Chinese Syntactic Analysis. Proceedings of 16th Workshop on Chinese Lexical Semantics (CLSW2015), Beijing.

Song T., Peng W., Song J., et al. 2016. The Construction of Sentence-Based Diagrammatic Treebank. Proceedings of 17th Workshop on Chinese Lexical Semantics (CLSW 2016), Singapore.

Xue N., Xia F., Chiou F., et al. 2005. The Penn Chinese Treebank: Phrase structure annotation of a large corpus. Natural Language Engineering, 11(2):207-238.

Wang Y., Guo, J., Che, W., et al. 2016. Transition-based Chinese semantic dependency graph parsing. Proceedings of China National Conference on Chinese Computational Linguistics, Yantai, 12-24.

褚晓敏, 奚雪峰, 蒋峰等. 2020. 宏观篇章结构表示体系和语料建设. 软件学报, 31(2):321-343.


(c) 2024 中国中文信息学会计算语言学专业委员会

郭丽娟, 彭雪, 李正华等. 2019. 面向多领域多来源文本的汉语依存句法树库构建. 中文信息学报, 33(02): 34-42.

胡壮麟. 1994. 语篇的衔接与连贯. 上海: 上海外语教育出版社.

李斌, 闻媛, 宋丽等. 2017. 融合概念对齐信息的中文 AMR 语料库的构建. 中文信息学报, 31(6): 93-102.

黎锦熙. 2001. 新著国语文法. 北京: 商务印书馆.

黎锦熙, 刘世儒. 1959. 汉语语法教材（第二编 词类和构词法）. 北京: 商务印书馆.

黎锦熙, 刘世儒. 1962. 汉语语法教材（第三编 复式句和篇章结构）. 北京: 商务印书馆.

廖序东. 1986. 论句子的图解——句子结构分析结果的表示法. 汉语学习, (01).

卢露, 矫红岩, 李梦等. 2022. 基于篇章的汉语句法结构树库. 自动化学报, 48(12): 2911−2921.

吕冀平. 1984. 句法分析和句法教学.《中国语文》杂志社. 汉语析句方法讨论集. 上海: 上海教育出版社出版: 216-235.

邱立坤, 金澎, 王厚峰. 2015. 基于依存语法构建多视图汉语树库. 中文信息学报, 29(3):9-15.

彭炜明. 2012. 句本位语法数字化平台的建设和应用研究. 北京师范大学博士学位论文.

彭炜明. 2021. 句本位语法的中文信息处理理论与实践. 北京: 外语教学与研究出版社.

彭炜明, 宋继华, 王宁. 2014. 基于句式结构的汉语图解析句法设计. 计算机工程与应用, 50(6):11-18.

宋柔, 葛诗利, 尚英等. 2017. 面向文本信息处理的汉语句子和小句. 中文信息学报, 31(02): 18-24.

宋柔. 2022. 小句复合体的语法结构. 北京: 商务印书馆.

奚雪峰, 褚晓敏, 孙庆英等. 2017. 汉语篇章微观话题结构建模与语料库构建. 计算机研究与发展, 54(08): 1833-1852.

谢晨晖, 胡正升, 杨麟儿等. 2023. 句式结构树库的自动构建研究. 中文信息学报, 37(2): 15-25.

荀恩东. 2023. 自然语言结构计算——意合图理论与技术. 北京: 人民邮电出版社.

杨天心, 彭炜明, 宋继华. 2014. 基于句式结构的高效语法图解标注系统. 中文信息学报, 28(04):43-49.

张志公. 1956. 语法和语法教学——介绍“暂拟汉语教学语法系统”. 北京: 人民教育出版社.

赵敏, 彭炜明, 宋继华等. 2014. 面向图解树库的标注工具开发与优化. 中文信息学报, 28(06): 26-33.

周强. 2004. 汉语句法树库标注体系. 中文信息学报, 18(04):1-8.

CCL 2024

# 英语科技论文摘要语步结构语料库构建研究

**李洪政*　王若锦*　冯冲†✉　刘芳***
*北京理工大学外国语学院，语言工程与认知计算工信部重点实验室/ 北京市房山区良乡东路9号
†北京理工大学计算机学院/ 北京市海淀区中关村南大街5号
{lihongzheng,wangruojin,fengchong,liufang}@bit.edu.cn

## 摘要

语步结构是学术论文中的文本语篇单位，在语步分析、论文写作等方面具有重要价值。尽管关于学术论文的语步研究非常丰富，但语步标注数据资源仍然相对较少。本研究开发构建了一个英语科技论文摘要语步结构标注语料库，目前已标注近3.4万个语步结构，涵盖了自然语言处理、计算机视觉、通信工程、机械工程等学科领域，同时进行了标注数据统计和分析。语料库构建的第一阶段依靠人工标注形成高质量语料，在第二阶段也是主要阶段，采用了基于BERT的自动识别与标注模型，在保证标注质量的同时能够提升标注速度，扩大标注规模。本研究基于构建的语料库开展了不同学科领域摘要语步结构识别实验，对比了我们的模型与ChatGPT和Claude3等大语言模型的识别效果。结果显示我们的模型在各类语步识别上的F1指标均优于大语言模型，表明了模型的有效性。该语料库目前可公开获取使用，能够为科技论文信息抽取、英语写作智能批改等自然语言处理相关任务和学术用途英语等外语教学与研究等提供必要的数据资源，同时也能有效推动外语教育数字化转型。

**关键词：** 语步结构；语料库；研究论文；摘要

# Research on Construction of Corpus for Move Structures in Abstracts of English Scientific Research Articles

**Hongzheng Li*　Ruojin Wang *　Chong Feng†✉　Fang Liu***
*School of Foreign Languages, Key Laboratory of Language, Cognition and Computation Ministry of Industry and Information Technology, Beijing Institute of Technology / No.9, Liangxiang East Road, Fangshan District, Beijing
†School of Computer Science, Beijing Institute of Technology / No.5, Zhongguancun South St., Haidian District, Beijing
{lihongzheng,wangruojin,fengchong,liufang}@bit.edu.cn

## Abstract

Move structures are discourse units in research articles (RA) and is of great value in move analysis, essay writing, etc. Although there is abundant research on move structures in academic articles, there are still relatively few move annotation data resources. This research developed and constructed a specific corpus for annotating move structures in English RA abstracts. Currently, nearly 34,000 Move structures have been annotated, covering the fields of Natural Language Processing (NLP), Computer Vision (CV), Communication Engineering and Mechanical Engineering. We also presented annotation statistics and analysis. The first stage of corpus construction relies on manual annotation to form high-quality corpus data. In the second and main stage, an automatic recognition and annotation model based on BERT is adopted, which can improve the annotation speed and expand the annotation scale while ensuring the annotation quality. We conducted move structure recognition experiments

第二十三届中国计算语言学大会论文集，第841页-第852页，太原，中国，2024年7月25日至28日。
卷1：主会论文
(c) 2024 中国中文信息学会计算语言学专业委员会

based on the constructed corpus, and compared the performance of our model with large language models(LLM) including ChatGPT and Claude3. The experimental results show that the F1 scores of move structure recognition achieved by our model outperformed those of LLM, indicating the effectiveness of the proposed model. This corpus is currently publicly available and can provide necessary data resources for NLP related tasks such as scientific paper information extraction and English writing intelligent assistance, it is also beneficial to foreign language teaching and research such as English for Academic Purposes, it can effectively promote the digital transformation of foreign language education.

**Keywords:** Move structure , Corpus , Research article , Abstract

## 1 引言

语步(Move)是指具有具体交际功能和目的的文本切分单位和语义片段，用来表示特定意义，代表着学术论文各章节的表述结构与序列(Swales, 1990; Swales, 2004)。语步能够有效表征论文语篇的宏观结构，并为实现整个语篇体裁的总体交际目的服务，揭示论文语篇的目的性、步骤性和规约性(杨延宁and 邹航, 2023)。

科技学术论文特别是研究论文（Research Article, RA）是科技和学术交流的重要载体。在研究论文中，摘要是必不可少的组成部分，通过简洁、清晰而有说服力的段落传达整篇文章的核心思想和亮点。好的摘要应该包括组织良好和逻辑合理的语步结构，分别论述论文的研究背景、研究目的、研究方法、研究结论等基本信息，以显示文章的重要性和价值。

下面是一个NLP领域的英文研究论文摘要示例，其中清楚包含了几种常见的语步结构。

**例句1**：⟨**研究背景**⟩Previous works on cross-lingual NER are mostly based on label projection with pairwise texts or direct model transfer. ⟨**研究差距**⟩However, such methods either are not applicable if the labeled data in the source languages is unavailable, or do not leverage information contained in unlabeled data in the target language. ⟨**研究目的**⟩In this paper, we propose a teacher-student learning method to address such limitations. ⟨**研究方法**⟩NER models in the source languages are used as teachers to train a student model on unlabeled data in the target language. ⟨**研究方法**⟩The proposed method works for both single-source and multi-source cross-lingual NER. ⟨**研究结论**⟩Extensive experiments for 3 target languages on benchmark datasets well demonstrate that our method outperforms existing state-of-the-art methods for both single-source and multi-source cross-lingual NER.

语步的概念最初来自修辞学，最早由美国语言学家J.Swales提出。语步研究在专门用途英语（English for Specific Purposes, ESP）和学术用途英语(English for Academic Purposes, EAP)等领域具有比较久远的历史，在外语科技论文写作、语步分析等也具有重要的应用价值(Hyland, 2008; Moreno and Swales, 2018; Swales, 2019)。自从语步被提出以来，很多研究基于不同学科领域学术论文的不同部分如摘要、引言、方法、结论等内容针对语步、修辞结构和语言学特征等开展了比较详细的分析和讨论(Cotos et al., 2017; Lu et al., 2021; Alsharif, 2023)。近年来也有一些研究工作关注语步结构的检测与自动识别等(Alliheedi et al., 2019; 张鑫et al., 2024; 王末et al., 2020; 王东波et al., 2018; 杜新玉and 李宁, 2024; 丁良萍et al., 2019)。语步结构识别可以使文献中蕴含的知识显式地体现出来，提高知识利用效率，是内容抽取、文本摘要等任务应用的重要基础性工作(黄红et al., 2022)。但是之前的研究工作更多地只关注某一特定学科领域论文中特定部分的语步分析，或某个特定语步的识别，几乎没有跨学科对比研究和全面的语步结构研究；另一方面，目前面向学术写作的语步结构自动标注的研究相对较少，同时也缺乏针对科技研究论文摘要部分中的语步标注资源建设。这种不足为跨语言语步分析与识别、科技论文信息抽取等自然语言处理任务和计算机辅助语言学习（Computer-Assisted Language Learning, CALL）等应用场景带来了很多挑战和障碍。

为了解决这些问题，本研究构建了一个大规模、多领域的英语科技论文摘要语步标注语料库，当前已累计标注近3.4万个语步实例，涵盖了人工智能、通信工程和机械工程等学科领域。

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

该语料库已在Github上开源发布[0]。标注主要分为两个阶段：在前期初始阶段，通过人工标注形成一部分高质量的标注数据；在后期阶段也是目前标注的主要阶段，在高质量标注数据的基础上，我们训练了一个BERT自动标注模型，实现语步自动标注，然后人工进行干预，修正系统标注不准确或者错误的语步，在快速扩大标注数量和规模的同时也能够保证标注质量。随后开展了语步结构自动识别实验，对比了我们的系统模型与GPT-4和Claude 3等大语言模型在不同学科领域的语步识别效果，同时针对具体实例进行了比较详细的分析。实验结果显示，本研究的模型在不同类型语步结构识别的F1指标分数均明显高于大语言模型的分数，表明了模型的有效性以及语料库建设的价值和必要性。

## 2 相关研究

### 2.1 学术领域语料数据资源

本研究构建的语料库属于学术领域语料库。学术文本语料库包含学术期刊、会议论文集等发表的研究论文、摘要等学术写作材料。学术语料库的建设有利于推动科学进步，在很多应用场景中具有实用价值，受到国内外学术界和业界的广泛关注。近年来，学术领域数据集的数量显著增加，同时也支持开源获取。比较具有代表性的语料库包括：ACL OCL语料库(Rohatgi et al., 2023)、S2ORC(Lo et al., 2020)、ARC(Bird et al., 2008)和AAN(Radev et al., 2009)等基于ACL Anthology构建的语料数据资源。这些数据集通常包含全面的元数据和科学论文全文，能够为研究人员提供可访问且重要的数据资源。

### 2.2 语步分类及语步标注

语步研究历史久远。不同时期的很多学者都针对语步进行了分类研究。例如：Swales开创性地提出了CARS语步模型(Swales, 1990; Swales, 2004)和语步分类体系(Swales and Feak, 2009)，为后来的语步研究带来了深远影响，很多研究者在此基础上不断改进语步的理论研究，包括Teufel等人提出的Argumentative Zoning (AZ) scheme(Teufel et al., 1999)及其修正后的理论等(Teufel et al., 2009; Teufel, 2010)。

随着语步理论研究的发展，也出现了一些在这些理论指导下建设的语步结构数据资源。Alliheedi等人重点关注生物化学领域论文方法部分中的语义角色和修辞语步(Alliheedi et al., 2019)。有研究者构建了选自《应用语言学》期刊（Applied Linguistics）摘要的数据集(刘霞, 2016)；Viera等人(2020)通过构建摘要语料库，调查了在英语母语国家和非英语国家发表的研究论文摘要中的修辞。已有的语步结构语料库通常都是面向特定单一领域的，而且规模相对较小，目前仍然缺乏多领域大规模的摘要语步语料库。因此，有必要构建涵盖多个学科领域的论文摘要标注语料库。

## 3 语料库构建过程

### 3.1 理论基础

本研究构建语料库的理论基础主要借鉴了Hyland关于论文摘要语步的五分类理论(Hyland, 2000)。他认为，学术论文全文中的语步分类可能并不完全适合分析摘要部分的语步结构。经过对纯科学、应用科学、人文科学和社会科学等学科门类的论文摘要进行了集中细致的考察，他提出学术论文摘要语步五分类法:

研究介绍（Introduction）：论述研究说明，包括研究重要性、关键术语概念、研究差距等

研究目的（Purpose）：指出该研究的一般或者特别目的

研究方法（Method）：指明该研究的具体方法，包括数据、处理过程等

研究产出（Product）：讨论该研究的主要发现和结果

研究结论（Conclusion）：引出该研究的结论，包括研究重要性、研究不足、对未来研究的启示等

Hyland的分类法对本研究具有重要的借鉴意义。但根据仔细的前期调研和小规模的先导标注试验（Pilot Study），我们发现Hyland的分类无法完全适合特定领域的摘要语步，因此需要进一步的详细分类体系。例如下面例句的类似表达在很多包括NLP在内的人工智能领域论文中

[0]https://github.com/ljk1228/RAAMove


(c) 2024 中国中文信息学会计算语言学专业委员会

都非常常见，一般是开源论文研究的相关信息，通常可以认为是对整个研究社区的贡献和价值，因此无法标注出Hyland分类中的任何一种。

**例句2：** We release source code for our models and experiments at Github.

在Hyland分类的基础上，我们根据具体标注需求，对该分类方法进行了调整和补充，提出了表 1所示的八种基本的语步类型，并为每种语步设置了标注标签。

| 语步 | 含义 | 标签 |
|---|---|---|
| 背景（Background） | 陈述研究领域，并提供关于研究的相关信息。 | BAC |
| 差距（Gap） | 表明当前的研究差距，提出开展研究的积极理由。 | GAP |
| 目的（Purpose） | 表明研究目的、假设等，概述论文背后的意图。 | PUR |
| 方法（Method） | 提供研究的设计、程序、假设、方法、数据等信息。 | MTD |
| 结果（Result） | 说明研究的主要发现或结果或已完成的工作。 | RST |
| 结论（Conclusion） | 总结研究结果或将结果扩展到论文范围之外。 | CLN |
| 启示（Implication） | 得出尚未明确陈述的推论以及对后续研究的启发等。 | IMP |
| 贡献（Contribution） | 指出研究带来的理论和实践价值。 | CTN |

Table 1: 本研究设计的语步类型、含义及其标签

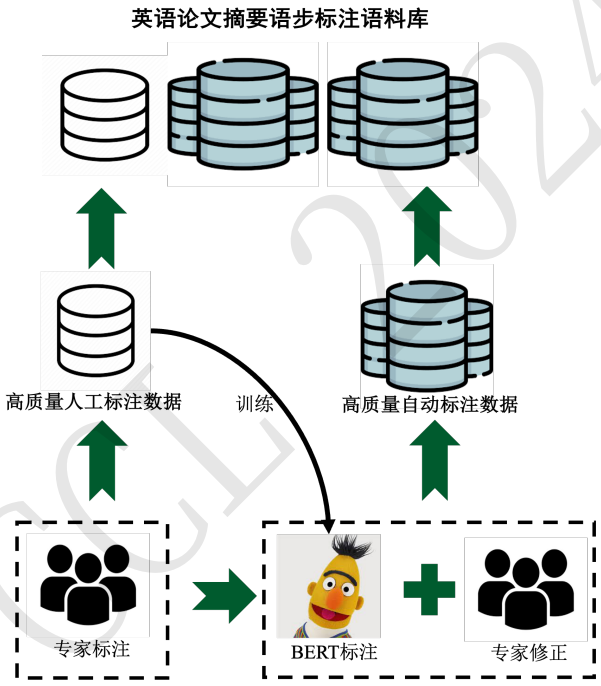



Figure 1: 语料库构建框架图

### 3.2 语料来源及处理

根据各专业领域的教授专家推荐，选择本领域的英文顶级期刊（JCR Q1）和国际知名顶级会议论文中的摘要作为标注语料，涵盖人工智能和工程学科大类，其中人工智能学科包括NLP和计算机视觉（CV）领域；工程大类包括通信工程（Communication Engineering, CE）和机械工程（Mechanical Engineering, ME）领域。

对于人工智能领域，我们选择了ACL大会主会论文和AAAI大会CV Track的会议论文。从ACL Anthology官网[1]可以直接下载收录了所有论文的BibTex文件，从中筛选出ACL2020-ACL2022三年的长论文摘要作为标注数据；另外补充了AAAI 2022大会CV Track的会议论文。

[1]https://aclanthology.org/


(c) 2024 中国中文信息学会计算语言学专业委员会

对于工程大类领域，选择了三本顶级期刊[2]。在Web of Science (WOS)检索平台中检索每本期刊的信息，然后将检索得到的论文题目、摘要等基本信息进行处理。为了方便标注语步结构，我们把两个学科领域的所有摘要文本段落根据主要的句末标点符号预处理为一句话一行的格式。同时在保留原文基本意义的前提下，对摘要文本进行必要的数据清洗，忽略可能会影响标注的特殊符号、引用格式等信息，以保证标注文本的质量。

### 3.3 语料库构建

如图1所示，语料库构建主要分为两个大的阶段：第一阶段是人工标注，第二阶段是系统自动标注+人工修正。我们采用开源标注工具Doccano[3]作为在线标注平台（图2）。

#### 3.3.1 人工标注

为了保证语步标注的质量和准确性，我们的语料标注团队由四名外国语学院的专业教师和一名外语语言学专业的博士生组成，同时每周举行例会讨论标注过程中存在的各种疑难问题。在标注过程中，原则上以完整的句子为单位，标注者需要为每个句子选择最合适的语步标签。

在标注平台中，选中一个句子以后即可弹出设计好的标注标签，标注者可以选择一个标签完成标注。如果需要修改，可以点击标签，选择其他合适的标签。

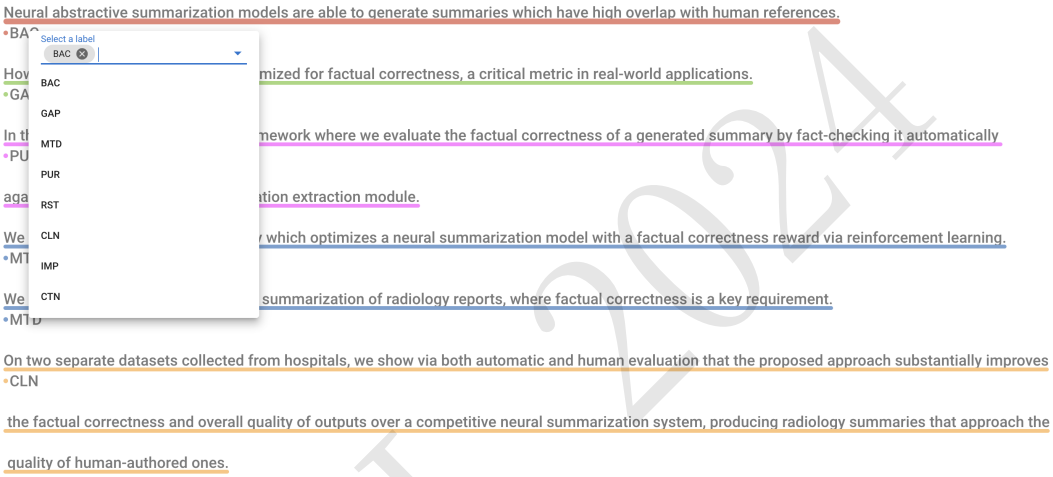



Figure 2: 标注平台界面

#### 3.3.2 自动标注

人工标注虽然能够保证标注质量，但无法提升标注速度和语料库标注规模，因此需要自动标注。本研究采用了Lin等人(2023)提出的基于BERT(Devlin et al., 2019)的语步自动标注模型(图 3)，将语步识别标注视为多标签识别和分类问题。

需要说明的是，在标注语步的时候，句中的不同单词对于预测句子语步类型具有不同的影响。像图3中的例句，句中有“results”“performance”等非常明显的关键词，它们对于快速识别为“RST”语步类型具有更高的贡献程度。考虑到这种普遍情况，该模型引入显著性注意力（saliency attention），句子中的每个单词都被视为一个特征，并计算其对特定语步类型的贡献（显著性值）。模型从人工标注的高质量数据中学习句子的语义特征，每个句子首先被分配一个表达整体语义的语步标签，然后设计了词语显著性向量（word saliency embeddings）与BERT的其他三种向量（即token、segment与position embeddings）一起作为输入表示来捕捉词语对于语步的贡献情况，提升模型识别语步的能力，最终完成语步识别和标注。

自动标注结果上传到标注平台以后，标注团队开始检查标注结果，修正标注不准确或者错误的语步结构。标注模型根据人工反馈的数据，可以不断进行迭代优化，进一步提升标注效

[2]三本期刊分别是机械工程领域的Journal of Mechanical Design、International Journal of Heat and Mass Transfer，通信工程领域的IEEE Journal on Selected Areas in Communications

[3]https://github.com/doccano/doccano


(c) 2024 中国中文信息学会计算语言学专业委员会

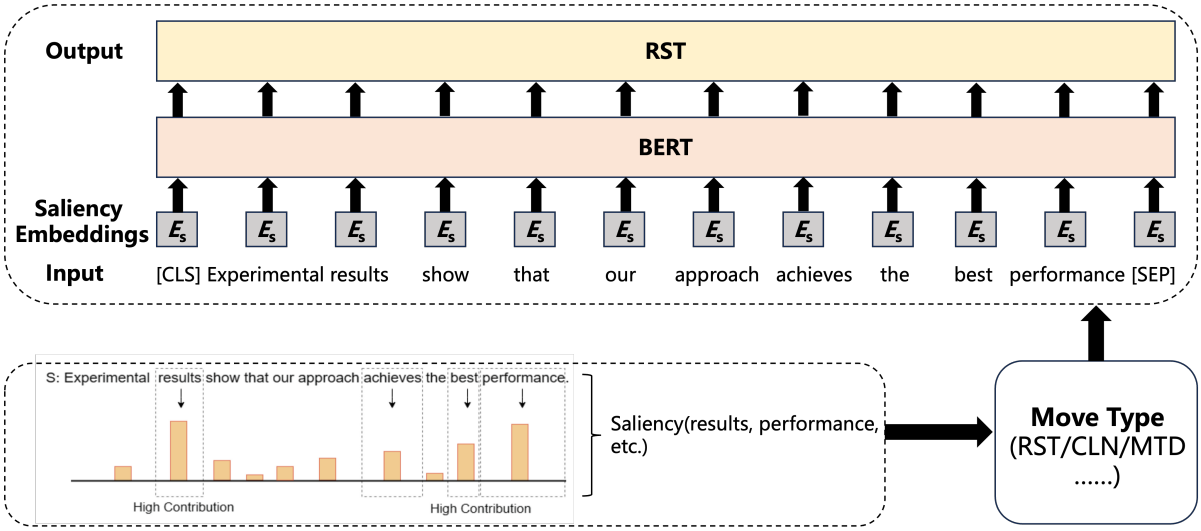



Figure 3: 语步自动识别模型架构

果。通过这种方式，可以快速提高标注速度和效率，在保证标注质量的同时扩大标注规模。语料标注完成以后，可以直接从标注平台中导出如下面实例所示的JSONL格式标注文件。

```
{ "id": 20, "data": "Words can have multiple senses. Compositional distributional models of meaning have been argued to deal well with finer shades of meaning variation known as polysemy, but are not so well equipped to handle word senses that are etymologically unrelated, or homonymy.", "label": [[0, 31, "BAC"], [32, 265, "GAP"]] }
```

## 4 语料库数据统计

本研究目前已标注2,670篇人工智能领域（NLP方向1,340篇，CV方向1,330篇）和2000篇工程大类（通信工程和信息工程各1,000篇）的论文摘要，共得到33,988个语步结构。两个学科领域论文摘要文本中的数据统计信息如表 2所示。其中，平均句子数、平均词语数和平均语步数分别指每篇摘要中平均包含的句子数量、单词数量和语步数量。

| | 摘要总数 | 句子总数 | 词语总数 | 平均句子数 | 平均词语数 | 平均语步数 |
|---|---|---|---|---|---|---|
| 人工智能领域 | 2,670 | 17,391 | 381,734 | 7 | 143 | 4 |
| 工程领域 | 2,000 | 16,597 | 406,244 | 8 | 202 | 4 |

Table 2: 两个学科领域的摘要文本数据统计

| 语步 | 人工智能 | 工程领域 | 小计 | % |
|---|---|---|---|---|
| BAC | 3,301 | 3,165 | 6,466 | 19.02 |
| GAP | 2,031 | 1,241 | 3,272 | 9.63 |
| PUR | 2,694 | 2,180 | 4,874 | 14.34 |
| MTD | **4,972** | **6,554** | **11,526** | **33.91** |
| RST | 2,181 | 1,551 | 3,732 | 10.98 |
| CLN | 1,494 | 1,512 | 3,006 | 8.84 |
| CTN | 601 | 229 | 830 | 2.44 |
| IMP | 117 | 165 | 282 | 0.83 |
| 总计 | 17,391 | 16,597 | 33,988 | 100 |

Table 3: 语料库中各类语步结构的标注和分布情况

从表 3可以看到，两个领域中标注数量前三位的语步依次均为MTD、BAC、PUR。不同


(c) 2024 中国中文信息学会计算语言学专业委员会

类型的语步分布存在很大差异："MTD"语步在两个学科中的标注次数都是最多的，两个领域的标注总量占比高达约34%，甚至超过了第二位"BAC"（6,466）和第三位"PUR"（4,874）的总和；而"IMP"的次数最少，只占到低于1%的比例。这表明不同领域的论文摘要均更关注MTD这一语步。这也跟直觉判断是一致的，因为方法确实是摘要中最有吸引力和最重要的部分之一，在摘要中通常会重点说明论文研究使用的方法，体现出该论文的价值和亮点工作。在标注过程中，我们也发现，很多论文摘要中的方法语步包括不止一个句子。

| | |
|---|---|
| NLP | [**BAC**]Knowledge graph (KG) entity typing aims at inferring possible missing entity type instances in KG. [**GAP**]It is a very significant but still under-explored subtask of knowledge graph completion. [**PUR**]In this paper, we propose a novel approach for KG entity typing which is trained by jointly utilizing local typing knowledge from existing entity type assertions and global triple knowledge in KGs. [**MTD**]Specifically, we present two distinct knowledge-driven effective mechanisms of entity type inference. ... [**CLN**]Experimental results on two real-world datasets (Freebase and YAGO) demonstrate the effectiveness of our proposed mechanisms and models for improving KG entity typing. [**CTN**]The source code and data of this paper can be obtained from Github... |
| ME | [**PUR**]This paper proposes a novel density-based method for structural design considering restrictions of multi-axis machining processes. [**MTD**]A new mathematical formulation based on Heaviside function is presented to transform the design field into a geometry which can be manufactured by multi-axis machining process. [**MTD**]The formulation is developed for 5-axis machining, which can be also applied to 2.5D milling restriction. The filter techniques are incorporated to effectively control the minimum size of void region. [**MTD**]The filter techniques are incorporated to effectively control the minimum size of void region. [**CLN**]The proposed method is demonstrated by solving the compliance minimization problem for different machinable freeform designs. |

Table 4: NLP领域和ME领域摘要语步标注实例对比

| | 人工智能领域 | | 工程领域 | |
|---|---|---|---|---|
| 语步类型 | #摘要 | 占比（%） | #摘要 | 占比（%） |
| BAC | 2,003 | 75.02 | 1,528 | 76.29 |
| GAP | 1,518 | 56.85 | 891 | 44.48 |
| PUR | **2,333** | **87.38** | **1,901** | **94.91** |
| MTD | 2,245 | 84.08 | 1,873 | 93.51 |
| RST | 1,540 | 57.68 | 953 | 47.58 |
| CLN | 1,192 | 44.64 | 1,079 | 53.87 |
| IMP | 112 | 4.19 | 159 | 7.94 |
| CTN | 544 | 20.37 | 215 | 10.73 |

Table 5: 两个学科领域中包括各类语步的摘要数量及占比情况

对比两个学科领域，可以看到不同学科的摘要文本特点和语步分布也存在较大差异。结合表2和表3，工程领域的摘要文本数量比人工智能少了近700篇，摘要文本句子总量整体也比人工智能领域少了近800句，但平均词语数量远高于人工智能，表明工程领域的摘要句子长度更长。语步分布上，工程领域的"MTD"语步数量远高于人工智能领域，其他语步如"BAC" "GAP" "PUR"等数量均少于人工智能。这表明通信工程和机械工程这两个工程学科的论文摘要更倾向用较多的语句来详细论述研究方法。标注团队在实践中也确实发现，很多论文摘要开头甚至很少交代研究背景和研究差距，而是直接用一句话交代研究目的，随即用多个连续的句子讨论研究采用了哪些方法，同时也不太注重提及研究本身的价值以及对于本领域的启发和贡献等。由于人工智能领域整体上具有明显的开源特点，很多论文摘要中经常提到研究


(c) 2024 中国中文信息学会计算语言学专业委员会

相关的代码、数据等面向公众开源，体现了对于研究社区的贡献，我们会把这种表述统一标注为“CTN”，因此这一语步在人工智能领域的数量更多。表 4通过两个标注实例对比了NLP和机械工程（ME）两个不同领域的摘要语步特点。

在语步分布的基础上，如果不考虑每种语步在摘要中的标注次数（一次或者多次），只要标注了该语步，则认为该摘要文本中包括该语步类型。我们也对这种情况进行了统计。表 5 显示了在2,670篇人工智能论文摘要和2,000篇工程领域论文摘要中，标注每种语步类型的摘要数量以及占全部摘要的比重。标注“PUR”的摘要数量最多，超过2,300篇的人工智能论文摘要中都有该语步，工程领域论文也类似。出现数量第二多的是“MTD”。对比表3和表4中的“PUR”和“MTD”语步，在全部摘要中标注数量最多的“MTD”高于“PUR”，是因为在一篇摘要中，通常只有一个句子被标注为“PUR”，但可能有多个句子被标注为“MTD”。

## 5 实验及分析

本部分基于已构建的语料库开展了语步结构的自动识别实验，分别从人工智能学科领域和工程学科领域中各抽样选择50篇研究论文摘要作为开放测试集，对比我们的标注模型与官方的大语言模型GPT-4和Claude 3(Opus)的识别效果。实验采用F1值作为评价指标。

### 5.1 实验数据及设置

表 6是实验数据的基本统计信息。平均句子数和平均句长分别是每篇摘要中平均包含的句子数量和句子中的单词数量。

| | 句子总数 | 平均句子数 | 平均句长 | 词语总数 | 词汇量 | 语步总数 |
|---|---|---|---|---|---|---|
| 人工智能领域 | 318 | 6.36 | 27.76 | 8,884 | 1,902 | 318 |
| 工程领域 | 397 | 7.94 | 32.12 | 12,752 | 2,415 | 397 |

Table 6: 数据集统计

为了尽可能地提高LLM识别语步类型的准确性，我们设计了下面的Prompt，并分别将其输入至官网的对话框中。大模型根据prompt给出反馈以后，正式开始语步识别任务。每个完整的摘要文本在两个模型中均只标注一次以确保标注公平。

“Move structures are important semantic and discourse units in research articles(RA). You are a senior expert in the field of EAP and are very good at analyzing the move structures in English RAs from different disciplines. You will analyze the move structures with the pre-defined move types and their labels as follows:

(1) Background(BAC): States the research area and provides any historical, theoretical, or empirical related information. (2) Gap(GAP): Establishes a niche: indicates a gap, adds to what is known, presents positive justification. (3) Purpose(PUR): Indicates purpose, thesis or hypothesis, outlines the intention behind the paper. (4) Method(MTD): Provides information on design, procedures, assumptions, approach, data, etc. (5) Result(RST): States main findings or results or what was accomplished. (6) Conclusion(CLN): Summarizes the results or extends results beyond scope of paper. (7) Implication(IMP): Draws inferences which has not been explicitly stated. (8) Contribution(CTN): Points out the theoretical and practical value of the methods used in the articles.

Please identity the most suitable move type and annotate it for [each sentence] in the abstract texts. That is, every complete sentence [must] have a move label. Here is an example:

[BAC] Recent neural models for relation extraction with distant supervision alleviate the impact of irrelevant sentences in a bag by learning importance weights for the sentences. [GAP]Efforts thus far have focused on improving extraction accuracy but little is known about their explanability.”

### 5.2 实验结果及分析

本研究的自动标注模型与两个大语言模型识别各类语步的对比结果如表 7所示。其中第二列是经人工核实后，数据集里每种语步类型的标准正确数量。图 4是八种语步识别F1指标的均


(c) 2024 中国中文信息学会计算语言学专业委员会

| | | Ours | | | GPT-4 | | | Claude3 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 正确语步总数 | #识别 | #正确 | F1(%) | #识别 | #正确 | F1(%) | #识别 | #正确 | F1(%) |
| AI领域 | BAC(61) | 64 | 60 | **96.00** | 50 | 42 | 75.68 | 65 | 52 | 82.54 |
| | GAP(32) | 31 | 30 | 95.24 | 42 | 29 | 78.38 | 32 | 22 | 68.75 |
| | PUR(48) | 49 | 45 | 92.78 | 46 | 40 | 85.11 | 53 | 43 | **85.15** |
| | MTD(101) | 94 | 92 | 94.36 | 94 | 84 | **86.15** | 86 | 78 | 83.42 |
| | RST(25) | 36 | 25 | 81.97 | 45 | 21 | 60.00 | 65 | 23 | 51.11 |
| | CLN(45) | 34 | 34 | 86.08 | 29 | 20 | 54.05 | 4 | 3 | 12.24 |
| | IMP(0) | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| | CTN(6) | 9 | 6 | 80.00 | 0 | 0 | 0 | 15 | 6 | 57.14 |
| 工程领域 | BAC(52) | 67 | 51 | 85.71 | 54 | 33 | 62.26 | 56 | 41 | 75.93 |
| | GAP(32) | 31 | 30 | 95.24 | 24 | 16 | 57.14 | 33 | 26 | 80.00 |
| | PUR(53) | 65 | 51 | 86.44 | 49 | 42 | 82.35 | 72 | 50 | 80.00 |
| | MTD(182) | 162 | 156 | 90.70 | 157 | 140 | **82.60** | 143 | 138 | **84.92** |
| | RST(11) | 28 | 11 | 56.41 | 62 | 10 | 27.40 | 82 | 11 | 23.66 |
| | CLN(62) | 39 | 38 | 75.25 | 5 | 4 | 11.94 | 2 | 2 | 6.25 |
| | IMP(3) | 3 | 3 | **100** | 4 | 1 | 28.57 | 0 | 0 | 0 |
| | CTN(2) | 2 | 2 | **100** | 0 | 0 | 0 | 9 | 1 | 18.18 |

Table 7: 两个学科领域语步自动识别标注实验结果

值。

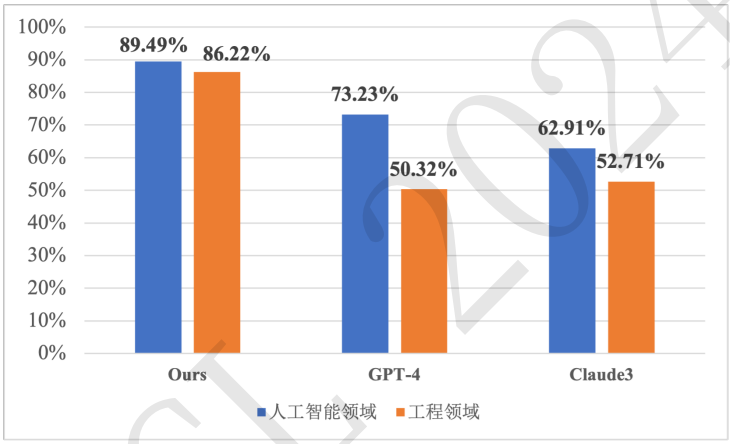



Figure 4: 三个模型整体识别所有语步的平均F1值

### 5.2.1 自动标注模型与大语言模型识别效果对比

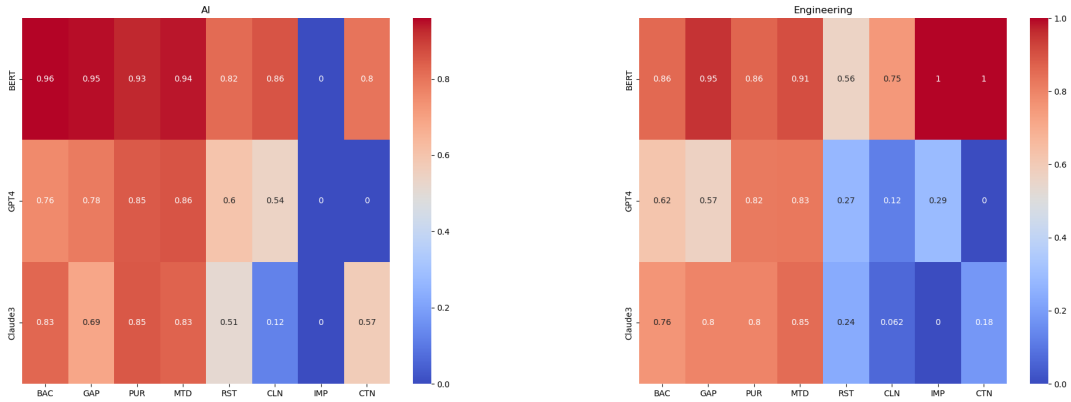
表 7显示，在当前实验中，本研究采用的基于BERT的识别模型在两个领域的语步识别F1指标均明显地远高于两个大语言模型的结果，而且识别每种语步类型的F1也相对更加稳定。

在人工智能领域和工程领域，我们的模型识别效果最好的F1值均在95%以上，所有语步整体识别的平均F1值也分别达到了89%和86%（图 4）；而GPT-4和Claude3的最高F1值仅接近85%，整体平均F1则低于70%。这表明了本文模型的有效性。但值得说明的是，并不是所有语句的识别效果都优于大模型。通过具体分析三个模型的识别结果，也发现了在某些摘要文本的语步识别中存在我们的模型识别错误、而大模型识别正确的情况。例如下面的实例：

**例句3：** [PUR]In this paper, we propose a novel bipartite flat-graph network (BiFlaG) for nested named entity recognition (NER), which contains two subgraph modules: a flat NER module for outermost entities and a graph module for all the entities located in inner layers. [MTD]Bidirectional LSTM (BiLSTM) and graph convolutional network (GCN) are adopted to jointly learn flat entities and their inner dependencies.

在该例句中，第二个句子出现在PUR语步之后，而且句中存在单词“adopted”，是比较明显的方法语步的信号。这个语步被我们的模型错误识别为BAC，但被GPT-4正确地识别


(c) 2024 中国中文信息学会计算语言学专业委员会




(a) 人工智能领域　　(b) 工程领域

Figure 5: 两个学科领域的语步识别F1指标热力图

为MTD。从中可以看出GPT对于文本语义的理解能力。

### 5.2.2 两个领域的识别效果对比

图 5是根据表 7绘制的语步识别F1指标热力图。从表 7和图 5中可以观察到，三个模型在AI领域的语步识别效果均明显优于工程领域。特别是工程领域，RST、CLN和CTN语步的识别效果都比较差。具体从两个领域来看，领域内各类语步的识别情况存在较大差异。在AI领域中，三个模型识别效果最好的语步类型均不相同，分别是BAC（我们的模型）、MTD（GPT-4）和PUR（Claude3）；在工程领域，我们的模型除了在IMP和CTN两个语步识别完全正确以外，识别最好的是GAP语步，其次是MTD，GPT4和Claude3则均在MTD上识别最好。这也再次说明了MTD语步在不同领域论文摘要中的重要性以及在语句表达上有其特殊之处，比其他语步更容易被识别。

### 5.2.3 大语言模型之间的识别效果对比

图 5显示，GPT-4在AI领域的识别效果明显好于Claude3，平均F1值比Claude3高了10个百分点；而在工程领域，Claude3的效果略好于GPT-4，但二者都在50%附近，远低于AI领域的平均F1，主要是受到了RST、CLN和CTN语步的影响。由此可见，整体上AI领域的摘要语步类型特征更加明显，更容易被识别出来。可以认为AI领域论文摘要写作的规范性和可读性好于工程领域。

尽管两个大语言模型的识别效果在两个学科领域中存在差异，但二者的识别表现也具有一定的趋同性。例如：在两个领域中，结论语步（CLN）的识别数量及准确识别数量均远远低于标准的正确数量，甚至出现了个位数的识别量，导致F1值急剧下降。我们试图跟GPT-4了解CLN识别过少的原因，它给出的解释是“CLN通常出现在章节或文档的末尾，如果提供的文本大部分来自文档的引言或方法部分，那么自然会导致CLN的实例减少。”换句话说，由于我们在Prompt中明确提到识别的是摘要文本中的语步，因此GPT-4认为摘要中不应该出现过多的CLN语步。

与CLN相反，两个领域中结果语步（RST）的识别数量则均超过了标准正确数量，也就是把大量原本不属于RST的语步识别为RST。这种情况在我们的识别模型中也很普遍。例如下面的实例：

**例句4：** [MTD]We propose a simple, effective transition-based model with generic neural encoding for discontinuous NER. [CLN]Through extensive experiments on three biomedical data sets, we show that our model can effectively recognize discontinuous mentions without sacrificing the accuracy on continuous mentions.

在该例句中，第二个句子的语步都被两个大模型识别为RST，但实际应该是CLN。因为如果句子中存在诸如“show”“demonstrate”之类的单词，一般就能比较明确地判断为CLN。


(c) 2024 中国中文信息学会计算语言学专业委员会

我们分析了GPT和Claude识别效果不稳定的原因，认为可能主要表现在两个方面：（1）我们只使用了官方网页版的大模型聊天模式，并未像专门训练我们的BERT模型那样有针对性地训练大模型，因此大模型无法充分学习到摘要文本中的语义信息。（2）我们在prompt中也没有为大模型提供太多的提示信息，包括能够快速且准确识别某种语步结构的关键信息等，例如“show”“demonstrate”等词语可以有助于快速判断为CLN语步。我们相信如果进一步优化prompt，将会提升大语言模型的语步识别效果。

## 6 结束语

本文面向英语科技论文写作，构建了一个涵盖多个学科领域、较大规模的论文摘要语步结构标注语料库。主要通过人工标注与自动标注+人工校对的方式构建，采用了基于BERT的语步自动识别与标注模型，能够保证语料库建设的规模和质量。在语步自动识别实验中，本文的模型在不同学科领域中的各类语步类型识别效果均优于大语言模型GPT4和Claude3的识别效果，体现了所构建的语料库和自动识别模型的有效性和价值。该语料库能够为科技论文信息抽取、科技论文智能辅助写作和批改等自然语言处理任务以及跨学科领域语步分析等外语教学和研究提供必要的数据资源支持，帮助二语学习者更好地理解论文语步结构，提升写作能力等，同时通过语言智能技术赋能外语场景，有助于推动外语教育数字化转型。

在未来的研究中，我们将在现有工作成果的基础上继续标注更多学科领域的摘要文本语步结构，持续扩大语料库的规模。我们同时也考虑标注科技论文中的其他组成部分，例如引言和方法部分，目标是建设大规模、多领域、多元化的语步数据资源，为开展科技论文写作智能批改等后续研究提供坚实基础。

## 致谢

本研究受到国家社科基金项目（23BYY166）资助。感谢项目课题组全体成员为本研究做出的贡献。感谢匿名评审为本文提出的宝贵意见和建议。

## 参考文献

Mohammed Alliheedi, Robert E. Mercer, and Robin Cohen. 2019. Annotation of rhetorical moves in biochemistry articles. In *Proceedings of the 6th Workshop on Argument Mining*, pages 113–123, Florence, Italy, August. Association for Computational Linguistics.

Maryam Alsharif. 2023. Rhetorical move structure in business management research article introductions. *Journal of Language and Linguistic Studies*, 18(4).

Steven Bird, Robert Dale, Bonnie Dorr, Bryan Gibson, Mark Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir Radev, and Yee Fan Tan. 2008. The ACL Anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may. European Language Resources Association (ELRA).

Elena Cotos, Sarah Huffman, and Stephanie Link. 2017. A move/step model for methods sections: Demonstrating rigour and credibility. *English for Specific Purposes*, 46:90–106.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, volume 1, pages 4171–4186.

Ken Hyland. 2000. *Disciplinary discourses: Social Interactions in academic writing*. Longman.

Ken Hyland. 2008. As can be seen: Lexical bundles and disciplinary variation. *English for specific purposes*, 27(1):4–21.

Jinkun Lin, Hongzheng Li, Chong Feng, Fang Liu, Ge Shi, Lei Lei, Xing Lv, Ruojin Wang, Yangguang Mei, and Lingnan Xu. 2023. Move structure recognition in scientific papers with saliency attribution. In *China Conference on Knowledge Graph and Semantic Computing*, pages 246–258. Springer.


(c) 2024 中国中文信息学会计算语言学专业委员会

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online, july. Association for Computational Linguistics (ACL).

Xiaofei Lu, Jungwan Yoon, and Olesya Kisselev. 2021. Matching phrase-frames to rhetorical moves in social science research article introductions. *English for Specific Purposes*, 61:63–83.

Ana I Moreno and John M Swales. 2018. Strengthening move analysis methodology towards bridging the function-form gap. *English for specific purposes*, 50:40–63.

Dragomir R. Radev, Pradeep Muthukrishnan, and Vahed Qazvinian. 2009. The ACL Anthology network corpus. In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries (NLPIR4DL)*, pages 54–61, Suntec City, Singapore, August. Association for Computational Linguistics (ACL).

Shaurya Rohatgi, Yanxia Qin, Benjamin Aw, Niranjana Unnithan, and Min-Yen Kan. 2023. The ACL OCL corpus: Advancing open science in computational linguistics. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10348–10361, Singapore. Association for Computational Linguistics.

John Swales and Christine B. Feak. 2009. *Abstracts and the writing of abstracts*. University of Michigan Press.

John Swales. 1990. *Genre analysis: English in academic and research settings*. Cambridge University Press.

John Swales. 2004. *Research genres: Explorations and applications*. Cambridge University Press.

John M Swales. 2019. The futures of eap genre studies: A personal viewpoint. *Journal of English for Academic Purposes*, 38:75–82.

Simone Teufel, Jean Carletta, and Marc Moens. 1999. An annotation scheme for discourse-level argumentation in research articles. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 110–117, Bergen, Norway, June. Association for Computational Linguistics (ACL).

Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2009. Towards domain-independent argumentative zoning: Evidence from chemistry and computational linguistics. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 1493–1502.

Simone Teufel. 2010. The structure of scientific articles: Applications to citation indexing and summarization. *CA: CSLI Publications*.

Rodrigo Tovar Viera. 2020. Rhetorical move structure in abstracts of research articles published in ecuadorian and american english-speaking contexts. *Arab World English Journal (AWEJ)*, 10.

丁良萍, 张智雄, and 刘欢. 2019. 影响支持向量机模型语步自动识别效果的因素研究. 数据分析与知识发现, 3(11):16–23.

刘霞. 2016. 英语学术论文摘要语步结构自动识别模型的构建. Ph.D. thesis, 北京外国语大学.

张鑫, 许海云, 杨宁, 方肖, and 赵爽. 2024. 有限样本下的科技文献语步识别方法探讨. 图书情报工作, 68:117–129.

杜新玉and 李宁. 2024. 中文学术论文全文语步识别研究. 数据分析与知识发现, 8:74–83.

杨延宁and 邹航. 2023. 基于语步结构的学术论文语篇构式研究. 外语教学理论与实践, 182(2):1–18.

王东波, 高瑞卿, 叶文豪, 周鑫, and 朱丹浩. 2018. 不同特征下的学术文本结构功能自动识别研究. 情报学报, 37:997–1008.

王末, 崔运鹏, 陈丽, and 李欢. 2020. 基于深度学习的学术论文语步结构分类方法研究. 数据分析与知识发现, 4:60–68.

黄红, 陈冲, and 张婧莹. 2022. 科技文献内容语义识别研究综述. 情报学报, 41(09):991–1002.


(c) 2024 中国中文信息学会计算语言学专业委员会

# Self-Guide:一种基于自我规划的大语言模型推理增强方法

刘艺彬[1], 刘正皓[1]*, 闫宇坤[2], 于是[2], 王硕[2], 杨麟儿[3], 陈慧敏[4], 谷峪[1], 于戈[1]
[1]东北大学计算机科学与工程学院，辽宁省沈阳市 110169
[2]清华大学计算机科学与技术系，北京市 100084
[3]北京语言大学信息科学学院，北京市 100083
[4]清华大学新闻与传播学院，北京市 100084
liuyibin@stumail.neu.edu.cn, {liuzhenghao, guyu, yuge}@mail.neu.edu.cn
{yanyk.thu, lineryang, wangshuo.thu}@gmail.com
{yus21, huimchen}@mails.tsinghua.edu.cn

## 摘要

尽管大语言模型在自然语言处理任务中取得显著进展，但其在复杂问题推理等领域还面临着认知负荷问题，即大语言模型在推理过程需要记忆并处理大量信息。因此，如何有效地减少语言模型推理过程中的认知负荷，缓解推理过程中可能出现的认知过载是一个亟待解决的问题。对此本文提出了Self-Guide方法，用于增强语言模型的推理能力。该方法通过指引大语言模型生成常识知识和推理指导，让语言模型基于自我规划来增强其推理能力，并通过与推理链结合的方式对模型的推理过程进行校准。与现有方法不同的是，本文在不对大语言模型进行微调或使用外部工具的情况下，显著提升了语言模型的推理性能。实验结果表明，Self-Guide方法在四种常见推理任务上性能显著优于基线方法，同时相比传统的推理链模型，Self-Guide方法在推理能力较弱的模型上也具有良好的泛化性能。通过结合大语言模型的自我规划和推理能力，Self-Guide方法为提升语言模型的推理能力提供了一种新的有效途径。

**关键词：** 大语言模型 ； 自我引导规划 ； 推理增强 ； 认知负荷

# Self-Guide: Enhancing LLM Reasoning Ability via Self-Plan

**Yibin Liu[1], Zhenghao Liu[1]*, Yukun Yan[2], Shi Yu[2], Shuo Wang[2], Liner Yang[3], Huimin Chen[4], Yu Gu[1], Ge Yu[1]**
[1]Department of Computer Science and Technology, Northeastern University, Shenyang, Liaoning
[2]Department of Computer Science and Technology, Tsinghua University, Beijing
[3]School of Information Science, Beijing Language and Culture University, Beijing
[4]School of Journalism and Communication, Tsinghua University, Beijing
liuyibin@stumail.neu.edu.cn, {liuzhenghao, guyu, yuge}@mail.neu.edu.cn
{yanyk.thu, lineryang, wangshuo.thu}@gmail.com
{yus21, huimchen}@mails.tsinghua.edu.cn

## Abstract

Despite significant advancements of Large language Models (LLMs) in NLP tasks, they still face the cognitive overload problem, especially in domains requiring complex reasoning. Specifically, during reasoning, LLMs need to process and memorize vast amounts of information during the reasoning process. Thus, it is a pressing issue to effectively reduce the cognitive load during the reasoning process of LLM to alleviate potential cognitive overload. We introduce the Self-Guide method to alleviate the problem, which boosts LLMs' reasoning abilities by leveraging self-generated common sense knowledge and reasoning instructions. Experimental results demonstrate that our Self-Guide method outperforms baseline methods significantly on four common reasoning tasks, which effectively enhances the reasoning ability of LLMs.

**Keywords:** LLMs , Self-Guide , Reasoning Enhancement , Cognitive Load

* 通讯作者

第二十三届中国计算语言学大会论文集，第853页-第869页，太原，中国，2024年7月25日至28日。
卷1：主会论文
(c) 2024 中国中文信息学会计算语言学专业委员会

## 1 引言

在自然语言处理及相关领域，大语言模型例如Chat-GPT、ChatGLM、MiniCPM、文心一言、通义千问(Brown et al., 2020; Du et al., 2021; Hu et al., 2024; Sun et al., 2021; Bai et al., 2023)等已经取得了重大进展。研究表明，随着模型参数规模的快速增长，语言模型展现出了一种涌现现象，使其在许多自然语言任务上的性能都显示了其有效性，同时在很多推理、规划任务中也展现了更为强大的能力(Wei et al., 2022a)。然而，尽管大语言模型在很多任务上表现出色，在解决复杂问题时候，其多步推理能力仍然面临着重要挑战(Valmeekam et al., 2022; Huang and Chang, 2022)。

传统的基于思维链（Chain-of-Thought, CoT）(Wei et al., 2022b)的推理方法在许多复杂的推理任务，如：数学计算、关系推理等任务中展现了良好的效果。其通过指令指引大语言模型对于给定问题进行一步步思考。在思维链的形成过程中，需要大语言模型在能够对问题的解决思路进行规划，同时又能够针对分解后的子问题进行回答，形成一个子问题-答案交融的推理链结果。澳大利亚新南威尔士大学的认知心理学家约翰·斯威勒（John Sweller）于1988年提出认知负荷理论（Cognitive Load Theory）(Sweller, 1988)。他认为人类的认知能力在处理新信息的过程中，其工作记忆的容量是相当有限的，因此我们需要考虑如何高效地呈现信息，来降低工作记忆的负荷量。考虑到语言模型的上下文学习能力和有限的上下文学习窗口(Dong et al., 2023)，在思维链生成过程中常常会面临的认知过载的问题，即产生幻觉或无效推理过程。因此，如何能够显著地降低大语言模型在思维链推理过程中所面临的认知负荷问题，缓解可能出现的认知过载是一个亟待解决的问题。

本文提出了Self-Guide方法——一种基于自我规划的大语言模型推理增强方法，用于增强语言模型的推理能力。该方法通过提示帮助大语言模型生成针对复杂问题的常识知识和推理指导（Guideline），让语言模型通过自我规划来预演其推理过程。并通过与推理链结合的方式来降低大语言模型在推理过程中所展现出的认知负荷。值得注意的是，我们在没有对大语言模型进行微调或使用外部工具的情况下，显著提升了语言模型在多种推理等任务上的性能表现。具体来说，本文的贡献主要包括以下三个方面：

- 本文提出了一种新的提示学习方法，通过大模型自身对给定问题的解决过程进行推演进而对给定问题产生相应的问题推理指导（Guideline）。
- 本文通过生成的指导在一定程度上缓解了传统基于思维链的推理模型在解决复杂问题中所具有的认知负荷问题，提高了语言模型的推理的性能和可靠性。
- 本文通过实验验证了较大规模模型通过规划所产生的问题推理指导能够有效地指导较小规模模型产生更加精准的结果，从而弥补了后者在推理任务上的不足。

我们的实验利用gpt-3.5-turbo-1106来实现Self-Guide方法，并在四种常见推理任务——语言理解、多跳问答、时间推理、关系推理上性能显著优于传统的基于思维链的推理模型（实现了平均约2%的提升）。同时我们通过实验证明了Self-Guide方法在推理能力较弱的模型（如LLaMA-7B-Chat和LLaMA-13B-Chat）上的泛化性，使其在自身推理能力较弱的条件下通过自我规划实现超过10%的提升。本文所有的数据和代码已在GitHub上开源。

## 2 相关工作

复杂推理为基于语言模型构建大量应用程序提供了机会，从而使语言模型有机会成为下一代“操作系统”(OpenAI et al., 2024)。因此大语言模型应当在复杂推理任务上具有较好的逻辑推理能力，并可以通过与工具、用户和外部环境的所有元素互动来完成复杂的指令。尽管大语言模型在很多任务上表现出色，多步推理能力仍然是其短板之一(Valmeekam et al., 2022; Huang and Chang, 2022)。为了增强大语言模型的推理能力，先前的研究提出了多种方法：Nye等人提出在生成最终答案之前，通过生成“Scratchpads”（即中间步骤）来微调语言模型，让语言模型具备合成/执行多步推理的能力(Nye et al., 2021)。推理链提示（Chain-of-Thought，CoT）

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版
项目资助：国家自然科学基金（No. 62106126，No. 62272093），辽宁省科技厅联合计划项目（No. 2023-MSBA-081）以及中央高校基本科研业务费项目（No. N2416012）。

以“Let’s think step by step”之类的指令来引导大语言模型先给出推理步骤以及中间的结果，通过推理链增强最终帮助大语言模型生成更加精准的结果。推理链模型在较大规模的语言模型上表现出优越的性能，然而在较小模型上其效果表现差强人意(Wei et al., 2022b)。此外，诱导语言模型生成推理原因需要构建大规模的数据集，或者仅使用少样本提示来牺牲准确性，因此Zelikman等人提出Self-Taught Reasoner(STaR), 利用大语言模型的现有推理能力，在给出答案之前迭代地生成高质量的推理原因(Zelikman et al., 2022)。Wang和Huang等人认为大语言模型有自我改进答案的能力，提出Self-Consistency解码策略，利用CoT生成多个推理路径和答案，最终选择出现次数最多的作为最终答案输出(Wang et al., 2022a; Huang et al., 2022)。

尽管使用高质量数据微调或少样本提示的方法显著增强了大语言模型的复杂推理能力，但在实际业务或问题中，常面临数据量不足或质量较低的情况，且微调和少样本提示的方法泛化性较低。为应对这一问题，Kojima等人证明大语言模型在Zero-Shot-CoT提示下仍然具有优秀的推理能力(Kojima et al., 2022)。而推理链提示在较难的问题上通常会出现推理与答案不一致的现象，为了克服从易到难的泛化挑战，Zhou等人提出Least-to-Most的提示方法，即将复杂问题分解为一系列更简单的子问题，然后按顺序解决它们(Zhou et al., 2022)。考虑到充分利用大语言模型海量的参数知识，Liu等人提出可以通过让语言模型生成知识作为额外输入来增强语言模型的常识推理能力(Liu et al., 2021)。

在大语言模型解决复杂问题时，语言模型的幻觉问题是另一个亟待解决的问题。由于预训练阶段记忆的带有噪音或错误的知识以及模型在特定任务上缺乏先验知识，大语言模型可能会出现生成文本不符合事实及不遵循原文的现象，即大语言模型的幻觉问题(Ji et al., 2023)。为缓解这一问题，研究者提出了多种方法：检索增强被证明可以显著减少幻觉问题。进一步地，Peng等人提出可以通过利用外部知识和语言模型自我反馈缓解大语言模型的幻觉(Peng et al., 2023)。同时，许多学者认为提示大语言模型在给出答案之前给出分步推理的过程能够缓解大语言模型的幻觉问题(Wei et al., 2023; Yao et al., 2022; Wang et al., 2022b; Li et al., 2022)，但这种假设只有在生成的推理符合模型真实推理的情况下才成立(Jacovi and Goldberg, 2020)。

然而语言模型生成的推理与其真实的推理过程并不总是保持一致(Turpin et al., 2024)。对此，Zhao等人提出Verify-and-Edit模型，其是一种知识增强的CoT提示框架，能够让语言模型利用外部知识编辑推理链来提高预测的真实性(Zhao et al., 2023)。Lyu等人提出Faithful CoT，通过将自然语言表示的推理过程转换为符号语言求解从而保证推理具有更高的忠实性(Lyu et al., 2023)。虽然，上述方法取得了较好的效果，但是在大语言模型进行推理时，传统方法往往需要语言模型同时处理问题规划和答案生成，可能导致认知过载的问题出现。为此，本文提出了Self-Guide模型，通过将语言模型的问题规划和答案生成进行解耦以期望提升大模型对复杂问题的推理效果。

## 3 Self-Guide：融合自身指导的大语言模型推理增强技术

在本节中，我们介绍了Self-Guide方法。传统的思维链增强推理方法（Chain-of-Thought, CoT）在具有较好的泛化性的前提下保持了良好的推理能力，但此方法在推理过程中需要同时对问题进行规划和回答，模型在整个推理过程需要处理并记忆大量信息，受到语言模型推理时间和自身能力的限制，语言模型通常会面临幻觉问题或无效推理，如何有效地减少思维链中的认知负荷，缓解大语言模型在推理过程中出现的认知过载问题是一个亟待解决的问题。

如图1所示，受到认知负荷理论(Sweller, 1988)的启发，本文提出了Self-Guide模型来缓解在传统推理链模型当中所面临的认知负荷问题。该方法融合了语言模型的自身指导对语言模型的推理过程进行增强，将语言模型的推理过程解耦为问题规划理解和答案生成。总的来说，我们首先介绍了基于大语言模型自适应问题规划方法（3.1节），然后我们利用语言模型的自适应性问题规划对其推理过程进行增强（3.2节）。

### 3.1 基于大语言模型自适应问题规划方法

对于给定的输入$Q$，Self-Guide模型通过问题规划生成以及基于问题规划的大语言模型增强两步方法提升大模型对于复杂问题的推理效果推理。该方法通过对传统推理链模型的问题规划以及答案生成进行解耦，以缓解大模型进行复杂推理过程中所具有的认知负荷问题，最终提升大模型的复杂推理效果。具体的提示模板在表5中展示。

根据认知负荷理论，我们采用Zero-Shot的设定，使大语言模型可以自适应地针对给定问题


(c) 2024 中国中文信息学会计算语言学专业委员会

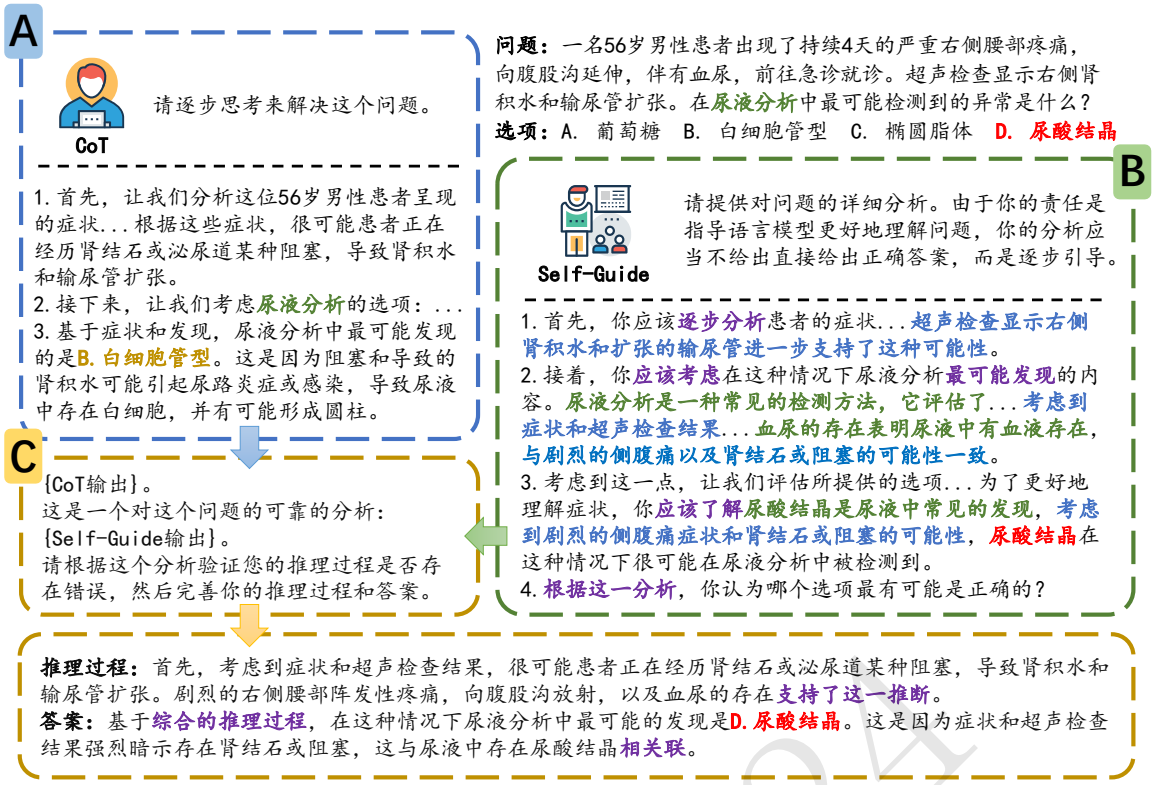



Figure 1: Self-Guide模型概述。模型分为思维链模块（Step A），基于大语言模型的自适应问题规划方法（Step B）和基于问题规划的大语言模型推理增强方法（Step C）。其中背景知识（Knowledge）、问题理解（Understanding）和问题规划（Planning）对应内容已高亮显示。

让大语言模型为问题的解决方案形成规划。对于给定的问题$Q$，语言模型生成规划Guideline。具体来说，它提供了基于大语言模型自身经验的指导方针，帮助语言模型减少推理过程过程中的无关认知负荷，从而将其注意力重新集中在推理过程上。对于给定问题$Q$，我们通过大语言模型生成任务解决规划：

$$\text{Guideline} = \text{Generate}(\text{Instruction\_Guide}, Q), \tag{1}$$

其中，Instruction_Guide表示语言模型结合问题$Q$生成自适应问题规划Guideline的过程，具体提示词模板见表5。通过这种方法利用大语言模型自适应地生成问题规划，以降低在复杂推理过程中可能出现的认知负荷，并提高大模型在复杂推理任务上的性能。

如图1中步骤B的输出部分所示，生成的Guideline包含了背景知识（Knowledge）部分作为对问题进行合理推理的基础，根据问题的特定领域提供相关的先验知识。对于大语言模型而言，它们已经通过海量的文本训练获取了广泛的知识，因此可以根据问题内容提供相关的背景知识，以帮助模型更好地理解和解决问题。问题理解（Understanding）部分需要模型准确把握问题的隐含信息、逻辑关系等，从而为后续的推理和解决提供基础。其关键在于从问题描述中提取有效信息，并将其与第一部分“背景知识”相结合，形成对问题的全面把握，从而有针对性地展开后续的推理和解决过程。问题规划（Planning）部分是Self-Guide中的关键内容，模型需要根据前面对问题的知识和理解，制定出一套合理的解决方案。通过合理的规划，模型可以避免盲目地进行推理，而是有条不紊地按照既定的步骤逐步推进，从而提高解决问题的准确性。

### 3.2 基于问题规划的大语言模型推理增强方法

根据Wei和Lanham等人的实验结果，规模较小的模型往往不具备根据链式推理方法提供推理过程的能力，如LLaMA-7B，LLaMA-13B等(Wei et al., 2022b)。，即在推理过程中表现出较高的内在认知负荷，当模型参数增大到大约10B时，大语言模型才会逐渐表现出推理能力(Lanham et al., 2023)。


(c) 2024 中国中文信息学会计算语言学专业委员会

由于GPT3.5模型有着优秀的基础推理能力，因此我们采取了一种自适应的学习流程（见图1）。首先，对于问题$Q$，我们让语言模型在独立思考和探索问题解决方案的过程中运用自身的认知资源，自主解决问题，并生成初步答案推理链：

$$\text{CoT} = \text{Generate}(\text{Instruction\_CoT}, Q), \tag{2}$$

其中推理链生成指令Instruction_CoT为“Let's think step by step”。接下来，我们让大语言模型利用公式1所产生的Guideline对语言模型的原始思维链进行增强。具体地，我们让大语言模型根据Guideline重新审视和反思初始答案CoT，修正推理过程中的错误，并根据Guideline完善推理过程和最终答案：

$$\text{GuideCoT} = \text{Template}(\text{CoT}, \text{Guideline}, Q), \tag{3}$$

其中模板如图1中步骤C的输入部分所示。最终，语言模型通过结合独立思考和自我规划的学习过程，生成信息密度更高且推理过程更缜密的答案。我们通过Self-Guide的自我规划在语言模型推理过程中减少了其在推理过程中冗余的推理步骤，从而有效降低了推理过程中的无关的认知负荷。这种方法旨在最大程度地发挥大语言模型的自主性和推理能力，并通过结合自我规划和自我反馈来提高针对于复杂推理问题的学习效果和答案的质量。

## 4 实验设计

在本节中，我们描述了实验中使用的数据集、基线、评估指标以及实现细节。

### 4.1 数据集：

在实验中，我们采用了知识理解、多跳问答、时间推理以及关系推理四个任务来评估Self-Guide模型在不同任务上的性能。由于ChatGPT的推理成本，对于数据集MMLU和CLUTRR，我们在实验中随机抽取了500个样本进行测试，这与之前的工作保持一致 (Yoran et al., 2023; Zhang et al., 2024; Xu et al., 2024)。由于StrategyQA和Date数据集的样本数量均小于500，因此我们在实验过程中在这两个数据集的全集上进行了实验。

**知识理解** 知识理解任务旨在衡量模型在预训练期间获得的知识。我们应用**MMLU**（大规模多任务语言理解）数据集(Hendrycks et al., 2021)，旨在通过对模型进行零样本或者少样本场景下的评估，来衡量预训练期间获得的知识。这使得该基准更具挑战性，也更类似于我们评估人类的方式。该基准涵盖57个学科，涉及科学、技术、工程和数学、人文科学、社会科学等。其难度从初级水平到高级专业水平不等，既测试世界知识，也测试解决问题的能力。这些科目的细化和广泛性使该基准成为识别模型知识薄弱领域的理想工具。

**多跳问答** 我们考虑了数据集**StrategyQA**(Geva et al., 2021)。这是一个开放域问题数据集，需要隐含的多步骤策略来回答，而回答问题所需的推理步骤隐含在问题中，应使用策略来推断。比如对于问题“亚里士多德使用笔记本电脑吗？”，回答时隐含着需要分别回答以下三个子问题“1.亚里士多德生活于何时？”“2.笔记本电脑是什么时候发明的？”“3.子问题2答案对应时间在子问题1答案对应时间的前面吗？”。

**时间推理** 我们采用来自**BIG-bench**基准测试(bench authors, 2023)的**Date Understanding**数据集。这是一个日期推理的数据集，旨在通过询问日期来测试语言模型理解日期的能力。它要求模型通过对相对时间段的计算，根据上下文推断日期。给定上下文，模型应该回答以下问题“What is the date [day] in MM/DD/YYYY?”，其中[day] 可以是today、tomorrow 等。虽然这项任务对人类来说很容易，但对于许多语言模型却很困难，因此用来评估各种优化模型推理的方法较有价值。

**关系推理** 我们使用**CLUTRR**基准(Sinha et al., 2019)，这是一个人物关系推理的数据集，要求自然语言理解（NLU）系统推断短篇小说中人物之间的亲属关系。要成功完成这项任务，既要提取实体之间的关系，又要推断出支配这些关系的逻辑规则。CLUTRR 允许通过评估逻辑规则的保留组合来精确测量模型的系统泛化能力，并允许通过添加经过策划的噪声事实来评估模型的鲁棒性。

### 4.2 基线

基于Zero-Shot的实验设定，我们将Self-Guide模型与不同的提示工程方法进行了对比。我们的基线主要由标准零样本提示（Zero-Shot）和链式推理方法（Chain-of-Thought，CoT）组成：


(c) 2024 中国中文信息学会计算语言学专业委员会

| 参数 | checkpoint | temperature | top_p | max_seq_len | max_gen_len | max_batch_size |
|---|---|---|---|---|---|---|
| 设置值 | meta_llama | 0.2 | 0.9 | 2048 | 128 | 8 |

Table 1: LLaMA大语言模型推理过程中的参数设置。

**标准零样本提示学习（Zero-Shot）：**在不提供示例的情况下，对于给定的问题，通过给定的指令让大语言模型直接输出预测答案。

**链式推理方法（Chain of Thought, CoT）：**在标准提示方法(Brown et al., 2020)的基础上，研究人员要求大语言模型在输出最终答案之前，显式输出逐步的中间推理步骤来增强大模型的常识和推理等能力。而这一系列逐步的中间推理步骤的内容就被称为链式推理方法(Wei et al., 2022b)。链式推理方法区别于传统的提示方法，实现了从输入到思维链再到输出的映射，即<Input-Reasoning Chain-Output>，而不是从输入直接到输出的映射<Input-Output>的方式。另外，链式推理方法分为Zero-Shot-CoT与Few-Shot-CoT。本实验采用Zero-Shot-CoT作为基线，不添加示例（即一些针对如何进行中间推理的示范内容），而仅在指令中添加“Let's think step by step”(Wei et al., 2022b)，就可以使大模型表现出较强的零样本推理能力。

### 4.3 评估指标

**准确率（Accuracy）：**我们使用准确率（Accuracy，Acc）来评估不同任务中模型的性能，这与先前的工作(Xu et al., 2024)类似。我们首先将大语言模型的输出和标准答案都转换为小写，然后执行字符串匹配（StringEM）方法来计算每个模型预测结果和标准答案之间的准确率。具体匹配规则见附录B。

**困惑度（Perplexity）：**对于Self-Guide方法在不同规模语言模型上的泛化能力实验，我们使用困惑度（Perplexity，PPL）来评估语言模型在推理过程中的生成文本质量。对于词元化的序列$X = (x_0, x_1, \ldots, x_t)$，序列$X$的困惑度为：

$$\mathrm{PPL}(X) = \exp\left\{-\frac{1}{t}\sum_{i}^{t}\log p_\theta(x_i|x_{<i})\right\}, \tag{4}$$

其中$\log p_\theta(x_i|x_{<i})$是根据我们的模型，第$i$个标记在前面的标记$x_{<i}$的条件下的对数似然。通常而言，PPL越小，表示文本的上下连贯度越高，文本越通顺。

### 4.4 实验细节

对于Self-Guide方法，即表2，表3以及表6，我们通过OpenAI API访问ChatGPT模型来评估其性能。具体地，我们使用gpt-3.5-turbo-1106作为推理的基础模型，并将temperature设置为0.2进行生成。对于Self-Guide方法在LLaMA模型上的泛化性实验，即表4和图2，我们首先使用gpt-3.5-turbo-1106作为自我规划的生成模型，然后分别在开源大语言模型LLaMA2-7B-Chat和LLaMA2-13B-Chat上评估生成的自我规划结果（Guideline）的性能，LLaMA模型的参数设定见表1。

## 5 实验结果

在本节中，我们展示了Self-Guide的整体性能。随后，我们通过消融实验证明了Self-Guide方法的有效性。接着，我们深入分析了该方法在较小规模模型上的泛化能力，Self-Guide方法的作用机制以及其机制的特征。最后，我们进行了案例分析。

### 5.1 Self-Guide模型整体性能效果

我们在表2中呈现了Self-Guide方法在不同领域数据集上的综合表现。在Zero-Shot的实验设置下，我们将Self-Guide方法与不同类型的基线进行了比较，包括标准Zero-Shot提示方法（Zero-Shot）和Zero-Shot设定下的链式思维提示方法（Chain of Thought, CoT）。

为了深入探究大语言模型如何通过自我规划来增强推理能力，我们采用了三种方法对大语言模型的推理进行增强：让模型直接通过自我规划进行推理（Self-Guide（No Instruct）），在


(c) 2024 中国中文信息学会计算语言学专业委员会

| 方法 | 多任务语言理解 MMLU | 多跳问答 StrategyQA | 时间推理 Date | 关系推理 CLUTRR |
|---|---|---|---|---|
| Zero-Shot | 56.0 | 68.2 | 54.0 | 36.6 |
| CoT | 66.2 | 72.7 | 64.9 | 55.6 |
| Self-Guide (No Instruct) | 65.8 | 69.8 | 61.6 | 47.4 |
| Self-Guide | 67.4 | 70.2 | 61.3 | 48.8 |
| **CoT w. Self-Guide** | **68.8** | **74.3** | **66.9** | **56.4** |

Table 2: 各模型实验结果。其中Self-Guide为我们的方法。在该实验中，所有方法均使用gpt-3.5-turbo-1106作为推理的基础模型。我们在附录展示了不同模型在MMLU数据集中不同领域上的实验结果（表6）。

| 方法 | 实验设置 | 多任务语言理解 MMLU | 多跳问答 StrategyQA | 时间推理 Date | 关系推理 CLUTRR |
|---|---|---|---|---|---|
| Zero-Shot | Vanilla | 56.0 | 68.2 | 54.0 | 36.6 |
| | w. Reflect | 56.6 | 47.1 | 51.0 | 24.0 |
| | w. Debate | 60.8 | 67.8 | 55.7 | 38.0 |
| | **w. Self-Guide** | **67.3** | **70.2** | **61.3** | **48.8** |
| CoT | Vanilla | 66.2 | 72.7 | 64.9 | 55.6 |
| | w. Reflect | 36.2 | 51.0 | 60.2 | 25.4 |
| | w. Debate | 63.2 | 65.3 | 61.3 | 36.4 |
| | **w. Self-Guide** | **68.8** | **74.3** | **66.9** | **56.4** |

Table 3: 消融实验结果。在该实验中，所有方法均使用gpt-3.5-turbo-1106作为推理的基础模型。其中Self-Reflect为在大语言模型生成答案后自我反思更新答案，Self-Debate为让大语言模型检查其他模型的答案后更新答案。

标准提示方法的基础上利用自我规划来增强推理过程并更新答案（Self-Guide），以及在链式推理方法下利用自我规划来增强推理过程并更新答案（CoT w. Self-Guide）。

如表2中实验结果所示，三种采用Self-Guide方法进行推理增强的方法相比基线模型（Zero-Shot）在所有数据集上都表现出较为显著的提升。特别值得注意的是，我们的模型CoT w. Self-Guide在所有数据集上均表现出最佳的性能。相较于Zero-Shot模型，我们的模型在所有数据集上的平均准确率都有显著提升（+6.1至+19.8），而相较于基于推理链的推理增强模型，我们的模型在所有数据集上的平均准确率也有一定程度的提升（+0.8至+2.6），证明了我们模型在解决复杂推理问题时的有效性，同时说明Self-Guide方法能够通过自然语言表达语言模型的所形成的自动规划，来缓解传统推理链中的认知负荷问题。

鉴于ChatGPT-3.5模型本身具备较强的推理能力，因此，仅仅使用Self-Guide方法进行自我规划（Self-Guide和Self-Guide（No Instruct））的效果相较于推理链模型，其并没有表现出显著提升。换言之，虽然Self-Guide模型能够通过生成Guideline的方式缓解大模型的认知负荷问题，但是诸如推理链模型当中针对于每一步规划生成相对应的中间结果亦是至关重要的。

### 5.2 消融实验

为了进一步验证Self-Guide方法的有效性，即Self-Guide所带来的性能提升是否是由于模型通过自我规划提高了推理能力，而不是由于模型的二次推理调用所产生更好地结果，我们进行了消融实验。

如表3中所示，我们将我们的模型与不同的推理增强方法进行比较。我们以标准提示方法（Zero-Shot）和链式思维推理方法（CoT）为基础答案（Vanilla），并使用不同的方法对基础答案进行增强，包括自我反思（Reflect）和对其他语言模型的推理过程进行检查（Debate）。

如实验结果所示，相比基线方法，让大语言模型对自身推理进行自我反思（Zero-Shot w. Reflect）时，相比较Zero-Shot模型，模型的推理能力略有下降（-21.1到+0.6）。而相比较推理链模型（CoT）时，自我反思（CoT w. Reflect）模型的推理能力下降的更为明显（-30.2到-


(c) 2024 中国中文信息学会计算语言学专业委员会

| 方法 | 实验设置 | 多任务语言理解 MMLU | 多跳问答 StrategyQA | 时间推理 Date | 关系推理 CLUTRR |
|---|---|---|---|---|---|
| **LLaMA2-7B-Chat** | | | | | |
| LLaMA-Guide | Zero-Shot | 44.8 | 59.6 | 28.4 | 27.2 |
| | CoT | 40.8 | 60.4 | 32.0 | 34.6 |
| GPT-Guide | CoT | 52.0 | 64.7 | 52.9 | 37.6 |
| | Guideline | 48.8 | **66.1** | 44.3 | 45.0 |
| | **CoT w. Guideline** | **53.0** | 65.5 | **54.9** | **45.6** |
| **LLaMA2-13B-Chat** | | | | | |
| LLaMA-Guide | Zero-Shot | 46.8 | 62.2 | 39.0 | 34.4 |
| | CoT | 45.6 | 51.2 | 45.7 | 39.4 |
| GPT-Guide | CoT | 54.0 | 62.7 | 59.9 | 43.0 |
| | Guideline | 54.2 | **65.9** | 57.4 | 46.0 |
| | **CoT w. Guideline** | **56.8** | 64.9 | **61.6** | **47.2** |

Table 4: Self-Guide在不同规模语言模型上的泛化能力实验结果。其中LLaMA-Guide方法由LLaMA模型独立完成推理问题，GPT-Guide方法由gpt-3.5-turbo-1106生成CoT和Guideline，在去除最终答案后指导LLaMA模型解决复杂的推理问题。

4.7）。而让语言模型对其他语言模型的推理过程进行检查并给出自身答案（Debate）时，相比较Zero-Shot模型，语言模型的答案准确率有所提升（-0.4到+4.8），但相比较CoT模型，CoT w. Debate模型的性能反而下降（-19.2到-3.0）。

实验结果表明，相比Reflect和Debate的推理增强方法，Self-Guide方法能够通过缓解大语言模型的认知负荷问题来提升了大语言模型的推理能力，而不是仅仅通过自我反思（Reflect）以及自我检查（Debate）以期望实现更加精准的推理过程。相比较Zero-Shot模型，Self-Guide方法使得语言模型在所有数据集上的推理能力提升了+2.0到+12.2，甚至在部分数据集中达到了和推理链模型相匹敌的效果。这证明了类似于推理链等诸多模型，其自我规划的结果能够在处理逻辑推理问题时起到至关重要的作用。

此外，Self-Guide也是一种有效方法能够通过与推理链结合的方式来进一步提升模型效果，Self-Guide方法使得语言模型在所有数据集上的推理能力提升了+0.8到+2.6。这表明，使用Self-Guide方法所取得的提升并不是由于模型的二次推理或调用所产生的，而是由于模型通过自我规划来对思维链进行校准，进而提升了大语言模型针对复杂任务的推理能力。

### 5.3 Self-Guide在不同规模语言模型上的泛化能力

该实验展示了Self-Guide在不同模型上的泛化能力，如表4所示，我们利用Self-Guide方法增强推理能力较弱的大语言模型，如LLaMA2-7B-Chat和LLaMA2-13B-Chat。使这些模型展现出更强的推理能力。由于较小规模的LLaMA模型基础推理能力较弱，无法很好地生成推理链并解决复杂的推理任务。在本实验中，我们尝试将ChatGPT-3.5视为教师模型，通过Self-Guide方法中生成的CoT和Guildeline来指导小规模语言模型解决复杂的推理问题（GPT-Guide）。具体地，在GPT-Guide的方法中，我们首先利用ChatGPT-3.5生成推理链和Guildeline，在去除ChatGPT-3.5生成的推理链和Guideline中的最终答案后，分别采用ChatGPT-3.5生成的推理链引导LLaMA模型推理（CoT），采用ChatGPT-3.5生成的Guideline引导语言模型进行推理（Guideline）和同时使用ChatGPT-3.5生成的CoT和Guideline来引导LLaMA模型进行推理（CoT w. Guideline）。

如表4所示，CoT w. Guideline方法在所有任务和数据集上都表现出了显著的性能，特别是在多跳问答、时间推理和关系推理任务上。具体来说，与由LLaMA模型独立完成推理问题的方法比较（LLaMA-Guide），在LLaMA2-7B-Chat模型上，我们的方法相比于标准Zero-Shot提示方法，性能提升范围为+5.9到+26.5；相比于CoT提示方法，性能提升范围为+5.1到+22.9。而在LLaMA2-13B-Chat模型上，相比于标准Zero-Shot提示方法，性能提升范围为+2.7到+22.6；相比于CoT提示方法，性能提升范围为+7.8到+15.9。这些结果表明，通过使用较大规模模型基于Self-Guide方法生成的Guideline来引导较小规模模型，较大规模模型所获得的丰富语义表示和推理能力可以有效地传递给原本推理能力较弱的较小规模模型，从而弥补了后者在推理任务上的不足。


(c) 2024 中国中文信息学会计算语言学专业委员会

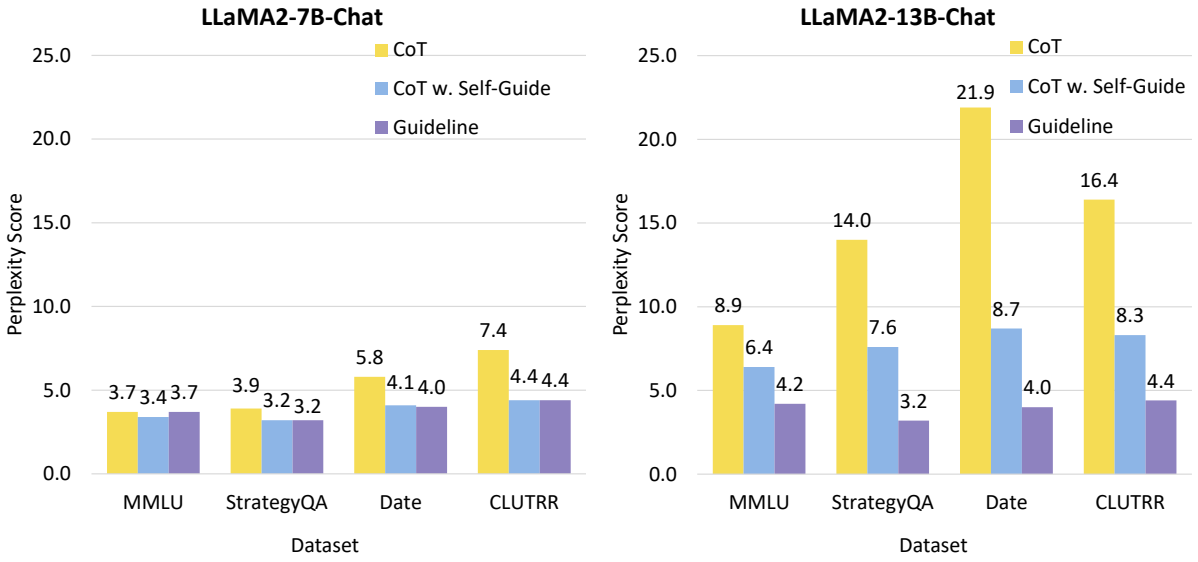



Figure 2: LLaMA模型在使用不同提示方法时困惑度（Perplexity）的差异。

除此之外，在使用ChatGPT-3.5模型的推理过程引导LLaMA完成推理问题的方法中（GPT-Guide），基于CoT w. Guideline引导的学习方法相比于基于CoT引导的学习方法，其推理性能都有一定程度的提升。在LLaMA2-7B-Chat模型上，CoT w. Guideline引导的模型相比于CoT引导的模型，性能提升范围为+0.8到+8.0。而在LLaMA2-13B-Chat模型上，CoT w. Guideline引导的模型相比于CoT引导的模型，性能提升范围为+1.7到+4.2。同时，随着模型参数的不断减小，CoT和CoT w. Guideline方法之间的性能差距也在不断扩大，从ChatGPT-3.5上的CoT和CoT w. Self-Guide之间的0.8%至2.6%，到LLaMA2-13B-Chat模型上CoT和CoT w. Guideline方法之间的1.7%至4.2%，到LLaMA2-7B-Chat模型上的CoT和CoT w. Guideline之间的0.8%至8.0%。这些结果表明由于模型的参数规模不同，其具有的认知容量也不同，推理链通过一边规划一边进行解答的过程所带来的认知负荷问题在小规模的语言模型中得到了放大。推理链中的问题解决思路难以适配到参数较小的模型当中。而Self-Guide模型所产生的Guildeline更类似于一种解题思路，其并未规定题目的具体解题方法，而是尝试引导模型考虑到解题当中的必要知识、步骤等，能够更好地适配较小规模语言模型的能力，以此帮助模型产生更好的生成结果。

为了更好地量化大语言模型在推理过程中的内在认知负荷，如图2所示，我们采用困惑度（Perplexity）来衡量语言模型在推理过程中生成答案的置信程度。如实验结果所示，相比起其他的提示方法，使用Guideline引导的模型在推理过程中困惑度显著降低，且使用LLaMA2-7B-Chat和LLaMA2-13B-Chat作为语言模型在所有数据集上PPL值稳定始终在4.0左右（-0.8到+0.4）。该现象证明了Self-Guide方法所产生的Guildeline能够更加容易地帮助不同规模的语言模型更加精确地产生正确答案，减小模型的不确定性。除此之外，Self-Guilde方法能够为大模型指导小模型学习提供一种可能。不同于通过推理链进行蒸馏的框架(Hsieh et al., 2023)，通过生成Guildeline进行蒸馏能够更加适配小模型的学习过程，帮助匹配较小规模语言模型的能力，进而能够更好地帮助其理解大模型问题解决思路、必要知识与规划，而不是一味地模仿更大规模模型的较难的解题思路。

### 5.4 案例分析

最后，我们在MMLU数据集中随机抽取了一个案例进行分析，以显示Self-Guide方法的有效性。其他数据集中的案例分析详见附录D。

如表5所示，经过Self-Guide方法增强后的模型的推理过程相比较推理链模型，其更加简洁且信息密度更高，同时推理步骤之间的逻辑关系也更加缜密。具体来说，在我们选择的案例中，Self-Guide首先定位问题领域（医学临床诊断），结合问题所述症状提供了与超声检查、尿液分析等相关的先验知识；之后提取问题内容（症状）的关键信息（血尿、右侧肾积水等），


（c）2024 中国中文信息学会计算语言学专业委员会

**问题：一名56岁的男子来到急诊科就诊，原因是他有4天的右侧腹部绞痛病史，疼痛还放射到腹股沟并伴有血尿。肾脏超声波检查显示右侧肾积水和输尿管扩张。尿液分析最有可能发现以下哪项?**
**选项：** A: 葡萄糖　　B: 白细胞管型　　C: 椭圆形脂肪体　　D: 尿酸结晶

**原始推理链（CoT）输入提示**
这里有一个问题：{问题} 为了解决这个问题，请逐步分析和推理，并给出你解决问题时的思考和推理过程。

**原始推理链（CoT）输出回答**
**1.** 首先，让我们分析一下这名56 岁男子的症状.....根据这些症状，患者很可能患有肾结石或某种形式的尿路梗阻，导致肾积水和输尿管扩张。
**2.** 现在，让我们考虑一下尿液分析的选项：A.....B.....C.....D.....
**3.** 根据症状和检查结果，最有可能的答案是B.白细胞管型，因为.....

**步骤1：基于Self-Guide的自我规划**

**获取Self-Guide的输入提示**
你是一位知识渊博、耐心细致的教授，其职责是指导学生正确解决问题。以下是一个问题：{问题} 注意：因为你的责任是引导学生回答问题，所以你的分析应该循序渐进，请注意你的角色是引导他们一步一步地解决问题，所以请不要给他们最终的结果。

**得到的Guideline**
**1.** 好吧，让我们一步一步解决问题。首先，我们来看看病人的症状...超声波检查显示右侧肾积水和输尿管扩张进一步证实了这种可能性。
**2.**现在，让我们来想想在这种情况下，尿液分析最有可能发现什么。尿液分析是一种常见的检查方法，用于评估尿液中的各种成分，包括血液、蛋白质、葡萄糖以及不同类型的细胞和结晶。根据症状和超声波检查结果..... 血尿的出现表明尿液中含有血液，这与腹部绞痛以及肾结石或梗阻的可能性是一致的。
**3.** 考虑到这一点，让我们对所提供的选项进行评估：A.....B.....C.....D..... 而尿酸结晶是肾结石患者尿液中常见的结晶，尤其是那些由尿酸组成的结晶。考虑到腹部绞痛的症状以及肾结石或梗阻的可能性，尿酸结晶是本病例尿液分析中最有可能发现的。
**4.** 根据以上分析，您认为哪个选项最有可能是正确的?

**步骤2：最终答案生成**

**获取最终答案的输入提示**
{CoT输出}
这是一个对这个问题的可靠的分析：{Self-Guide输出}。
请根据这个分析验证您的推理过程是否存在错误，然后完善你的推理过程和答案。

**得到的最终答案**
**推理过程：** 首先，考虑到症状和超声波检查结果，患者很可能患有肾结石或某种形式的尿路梗阻，导致肾积水和输尿管扩张。右侧腹部绞痛、向腹股沟放射的疼痛和血尿也支持这一推断。接下来，让我们对尿液分析的选项进行评估：A. 尿液中的葡萄糖通常与糖尿病有关，在这种情况下是不可能出现的。B. 白细胞管型–虽然表明有炎症或感染，但在本病例中可能不是最有可能的发现。C. 椭圆形脂肪体–与患者的症状和超声检查结果没有直接关系。D. 尿酸盐结晶–考虑到患者有腹部绞痛的症状，以及肾结石或梗阻的可能性，尿酸盐结晶是本病例最有可能的尿液检查结果。
**最终答案：** 根据综合推理过程，本病例尿检最有可能发现的是尿酸结晶（选项D）。这是因为症状和超声波检查结果强烈提示存在肾结石或梗阻，而肾结石或梗阻与尿液中存在尿酸结晶有关。

Table 5: 案例分析。我们随机选择了一个MMLU数据集中的例子来展示Self-Guide的效果，并突出显示了答案和相关内容。这是一个经过翻译的中文版本。原始内容在表7中展示。

将这些和先验知识结合，形成对整个临床病情的深入理解；最后根据深入理解，提供了从各个选项开始的分析方案。在通过Self-Guide得到Guideline之后，模型能够首先清晰地给出针对于整个问题的推理过程，其结合了Guideline所提供的理解内容并按照Self-Guide从选项逐步剖析，并结合初始的推理链进行反思与分析，最后直接给出了正确答案。这表明Self-Guide充分利用了模型自身的强大生成能力，使其能够根据具体问题灵活地调整推理策略，从而更好地应对各种复杂推理场景。Self-Guide能够帮助大语言模型对思维链结果进行反思，引导大语言模型对问题进行细致的解析，在处理复杂推理问题时能够通过形成知识密集的Guildline的相关内容，以在处理复杂推理任务时降低大语言模型所面临的认知负荷问题。

## 6 结论

本文提出了一种结合了大语言模型生成的常识知识和推理指导，让语言模型基于自我规划来增强其推理能力的方法Self-Guide。在不对语言模型进行微调的同时显著增强了语言模型的推理能力。我们证明了我们的方法在四种常见推理任务——语言理解、多跳问答、时间推理、关系推理上性能显著优于基线。相关实验结果分析进一步说明了Self-Guide模型可以通过Guildeline将较大规模语言模型的语言表示和推理能力有效地传递给原来推理能力较弱的较小规模模型，同时显著地降低较小规模模型生成答案时候的不确信度。

## 参考文献

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report.

BIG bench authors. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. Transactions on Machine Learning Research.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A survey on in-context learning.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2021. Glm: General language model pretraining with autoregressive blank infilling. arXiv preprint arXiv:2103.10360.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding.

Cheng-Yu Hsieh, Chun-Liang Li, CHIH-KUAN YEH, Hootan Nakhost, Yasuhisa Fujii, Alex Jason Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In The 61st Annual Meeting Of The Association For Computational Linguistics.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zheng Leng Thai, Kaihuo Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. Minicpm: Unveiling the potential of small language models with scalable training strategies.

Jie Huang and Kevin Chen-Chuan Chang. 2022. Towards reasoning in large language models: A survey. arXiv preprint arXiv:2212.10403.

Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2022. Large language models can self-improve. arXiv preprint arXiv:2210.11610.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? arXiv preprint arXiv:2004.03685.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. ACM Comput. Surv., 55(12), mar.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. Advances in neural information processing systems, 35:22199–22213.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. 2023. Measuring faithfulness in chain-of-thought reasoning. arXiv preprint arXiv:2307.13702.

Shiyang Li, Jianshu Chen, Yelong Shen, Zhiyu Chen, Xinlu Zhang, Zekun Li, Hong Wang, Jing Qian, Baolin Peng, Yi Mao, Wenhu Chen, and Xifeng Yan. 2022. Explanations from large language models make small reasoners better.


(c) 2024 中国中文信息学会计算语言学专业委员会

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2021. Generated knowledge prompting for commonsense reasoning. arXiv preprint arXiv:2110.08387.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. arXiv preprint arXiv:2301.13379.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. 2021. Show your work: Scratchpads for intermediate computation with language models. arXiv preprint arXiv:2112.00114.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and Jianfeng Gao. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback.

Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L. Hamilton. 2019. CLUTRR: A diagnostic benchmark for inductive reasoning from text. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, Proceedings of the 2019 Conference on Empirical Methods in Natural


(c) 2024 中国中文信息学会计算语言学专业委员会

Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 4506–4515, Hong Kong, China, November. Association for Computational Linguistics.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, et al. 2021. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. arXiv preprint arXiv:2107.02137.

John Sweller. 1988. Cognitive load during problem solving: Effects on learning. Cognitive science, 12(2):257–285.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2024. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. Advances in Neural Information Processing Systems, 36.

Karthik Valmeekam, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. 2022. Large language models still can't plan (a benchmark for llms on planning and reasoning about change). arXiv preprint arXiv:2206.10498.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022a. Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. 2022b. Rationale-augmented ensembles in language models.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022a. Emergent abilities of large language models. arXiv preprint arXiv:2206.07682.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35:24824–24837.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Zhipeng Xu, Zhenghao Liu, Yibin Liu, Chenyan Xiong, Yukun Yan, Shuo Wang, Shi Yu, Zhiyuan Liu, and Ge Yu. 2024. Activerag: Revealing the treasures of knowledge via active learning.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. arXiv preprint arXiv:2210.03629.

Ori Yoran, Tomer Wolfson, Ben Bogin, Uri Katz, Daniel Deutch, and Jonathan Berant. 2023. Answering questions by meta-reasoning over multiple chains of thought.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. Star: Bootstrapping reasoning with reasoning. Advances in Neural Information Processing Systems, 35:15476–15488.

Jintian Zhang, Xin Xu, Ningyu Zhang, Ruibo Liu, Bryan Hooi, and Shumin Deng. 2024. Exploring collaboration mechanisms for llm agents: A social psychology view.

Ruochen Zhao, Xingxuan Li, Shafiq Joty, Chengwei Qin, and Lidong Bing. 2023. Verify-and-edit: A knowledge-enhanced chain-of-thought framework. arXiv preprint arXiv:2305.03268.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. 2022. Least-to-most prompting enables complex reasoning in large language models. arXiv preprint arXiv:2205.10625.

CCL 2024

| 方法 | 多任务语言理解(MMLU) | | | | |
|---|---|---|---|---|---|
| | 全部 | 人文伦理 | 社会科学 | 理工科 | 其他 |
| Zero-Shot | 56.0 | 47.2 | 67.8 | 46.3 | 68.3 |
| CoT | 66.2 | 59.0 | 74.8 | 63.4 | 71.3 |
| Self-Guide | 65.8 | 57.4 | 73.9 | 55.4 | **82.7** |
| Zero-Shot w. Self-Guide | 67.4 | **62.6** | 72.2 | 61.2 | 77.6 |
| **CoT w. Self-Guide** | **68.8** | 58.4 | **76.5** | **65.9** | 80.2 |

Table 6: 不同模型在MMLU数据集中不同领域上的实验结果。其中Self-Guide为我们的方法。在该实验中，所有方法均使用gpt-3.5-turbo-1106作为推理的基础模型。

# 附录

## A 模型在MMLU数据集中不同领域上的性能

MMLU基准涵盖57个学科，涉及科学、技术、工程和数学、人文科学、社会科学等。其难度从初级水平到高级专业水平不等，既测试世界知识，也测试解决问题的能力。这些科目的细化和广泛性使该基准成为识别模型知识薄弱领域的理想工具。我们在所有学科（All），人文伦理（Hum.），社会科学（Soc. Sci.），理工科（STEM），和其他问题（Other）等多个领域进行了实验（表6）。

## B 评估指标细节

在执行字符串匹配（StringEM）方法来计算每个模型预测结果和标准答案之间的准确率的过程中，具体地，我们首先将大语言模型的输出和标准答案都转换为小写。对于MMLU和StrategyQA数据集，我们提取大预言模型生成结果中的第一个单词作为模型预测结果与标准答案进行匹配。对于Date数据集，我们提取出所有由数字、斜杠、数字、斜杠、数字组成的模式串，在转换为标准日期格式（MM/DD/YYYY）后与标准答案进行匹配。对于CLUTRR数据集，我们直接执行字符串匹配方法来计算每个模型预测结果和标准答案之间的准确率。

## C 英文原版案例分析

我们在MMLU数据集中各随机选择了一个例子来展示Self-Guide的效果，并突出显示了答案和相关内容。我们在表5中展示了案例分析的原始输入和输出。为了方便阅读我们在正文中展示了简略版的中文案例分析。

## D 其他数据集案例分析

我们在CLUTRR和Date数据集中各随机选择了一个例子来展示Self-Guide的效果，并突出显示了答案和相关内容。由于StrategyQA数据集需提交至Leaderboard进行评测，我们无法知道单个例子的对错，因此不对StrategyQA进行案例分析。


(c) 2024 中国中文信息学会计算语言学专业委员会

**问题：A 56-year-old man comes to the emergency department because of a 4-day history of colicky right flank pain that radiates to the groin and hematuria. Ultrasound examination of the kidneys shows right-sided hydronephrosis and a dilated ureter. Which of the following is most likely to be found on urinalysis?**
选项：A: Glucose B: Leukocyte casts C: Oval fat bodies D: Uric acid crystals

**原始推理链（CoT）输入提示**
Here is a question:{Question} In order to solve this problem, analyze and reason step by step, and give your thinking and reasoning process while solving the problem.

**原始推理链（CoT）输出回答**
**1.** First, let's analyze the symptoms ... Based on these symptoms, it is likely that the patient is experiencing a kidney stone or obstruction in the urinary tract, leading to hydronephrosis and dilated ureter.
**2.** Now, let's consider the options for urinalysis:
A. Glucose in the urine is typically associated with diabetes and would not be expected in this case.
B. Leukocyte casts in the urine indicate inflammation or infection in the kidneys or urinary tract...
C. Oval fat bodies are associated with nephrotic syndrome and would not be expected in this case.
D. Uric acid crystals are associated with gout, but they are not associated with the symptoms ...
**3.** Based on the symptoms and findings, the most likely answer would be B.leukocyte casts, Because...

**步骤1：基于Self-Guide的自我规划**

**获取Self-Guide的输入提示**
You are a knowledgeable and patient professor whose role is to guide students in solving problems correctly. Here is a question: {Question} Note: Since your responsibility is to guide students in answering the question, your analysis should think step by step, Please note that your role is to guide them step by step through the problem, so please don't give them the final result.

**得到的Guideline**
**1.** Alright, let's break this down step by step. First, let's consider the symptoms presented by the patient. ... The ultrasound examination showing right-sided hydronephrosis and a dilated ureter further supports this possibility.
**2.** Now, let's think about what would be most likely to be found on urinalysis in this case. Urinalysis is a common test that evaluates various components of urine, including the presence of blood, protein, glucose, and different types of cells and crystals. Given the symptoms and the ultrasound findings, ... The presence of hematuria suggests that there is blood in the urine, which is consistent with the colicky flank pain and the possibility of a kidney stone or obstruction.
**3.** Considering this, let's evaluate the options provided:... Therefore, we can eliminate option D. uric acid crystals are a common finding in the urine of patients with kidney stones, especially those composed of uric acid. Given the symptoms of colicky flank pain and the possibility of a kidney stone or obstruction, uric acid crystals are the most likely finding on urinalysis in this case.
**4.** Based on this analysis, which option do you think is most likely correct?

**步骤2：最终答案生成**

**获取最终答案的输入提示**
{CoT-log}
This is another analysis of this question:{Self-Guide}. Please combine this reasoning process with your reasoning processes, then give me a better reasoning process.
Generation Format: 1. Inference process: 2. Answer:

**得到的最终答案**
**Inference process:**
**1.** First, considering the symptoms and ultrasound findings, it is likely that the patient is experiencing a kidney stone or some form of obstruction in the urinary tract, leading to hydronephrosis and dilated ureter. The presence of colicky right flank pain, radiating to the groin, and hematuria supports this inference.
**2.** Next, let's evaluate the options provided for urinalysis: A. Glucose in the urine is typically associated with diabetes and would not be expected in this case. B. Leukocyte casts - While indicative of inflammation or infection, it may not be the most likely finding in this case. C. Oval fat bodies - Not directly related to the symptoms and ultrasound findings presented by the patient. D. Uric acid crystals. Given the symptoms of colicky flank pain and the possibility of a kidney stone or obstruction, uric acid crystals are the most likely finding on urinalysis in this case.
**Answer:** Based on the combined reasoning processes, the most likely finding on urinalysis in this case is uric acid crystals (option D). This is because the symptoms and ultrasound findings strongly suggest the presence of a kidney stone or obstruction, which would be associated with the presence of uric acid crystals in the urine.

Table 7: 案例分析。我们随机选择了一个MMLU数据集中的例子来展示Self-Guide的效果，并突出显示了答案和相关内容。这是英文原始输入和输出内容，经过翻译的中文版本在正文表5中展示。


(c) 2024 中国中文信息学会计算语言学专业委员会

**信息**：[Davis] played chess with his brother [Rick]. [Davis] was disappointed that his father, [Kent], wouldn't be at the play to see him perform. [Sharon] likes taking her son [Oliver] to baseball practice. She also likes taking her son [Rick] to soccer practice. [Dorothy] made cookies with her grandson [Davis]. [Davis]'s father [Kent] is not allowed to eat sugar.
**问题**：**How is [Dorothy] related to [Oliver]?**
**答案**：grandmother

**原始推理链（CoT）输入提示**
To solve the problem, Please think and reason step by step, then answer. Question: {Question}
Generation Format: 1. Reasoning process: 2. Answer:

**原始推理链（CoT）输出回答**
**Reasoning process:**
**1.** [Davis] is the grandson of [Dorothy].
**2.** [Oliver] is the son of [Sharon].
**3.** There is no direct relationship mentioned between [Dorothy] and [Oliver].
**Answer:** There is no direct relationship between [Dorothy] and [Oliver] mentioned in the information.

**步骤1：基于Self-Guide的自我规划**

**获取Self-Guide的输入提示**
You are a knowledgeable and patient professor whose role is to guide students in solving problems correctly. Here is a question:{Question} please provide a detailed analysis. Note: Since your responsibility is to guide students in answering the question, your analysis should think step by step, Please note that your role is to guide them step by step through the problem, so please don't give them the final result.

**得到的Guideline**
**Step 1:** Identify the family members mentioned in the story:
- Davis - Rick - Kent - Sharon - Oliver - Dorothy
**Step 2:** Determine the relationships between the family members that are explicitly mentioned in the story: - Davis is the grandson of Dorothy - Rick is the brother of Davis - Kent is the father of Davis - Sharon is the mother of Oliver and Rick
**Step 3:** Identify any indirect relationships that may help us determine the relationship between Dorothy and Oliver: - Davis is the grandson of Dorothy - Kent is the father of Davis - Sharon is the mother of Oliver - Sharon is also mentioned as the mother of Rick
**Step 4:** Consider the relationships between the family members and how they might connect Dorothy and Oliver: - Since Davis is the grandson of Dorothy, and Kent is the father of Davis, it means that Kent is the son of Dorothy. This makes Dorothy the grandmother of Kent's children, including Oliver.
**Step 5:** Based on the relationships identified, we can infer that Dorothy is the grandmother of Oliver. Now, based on the analysis, we can determine the relationship between Dorothy and Oliver.

**步骤2：最终答案生成**

**获取最终答案的输入提示**
{CoT-log}
This is a credible analysis of this question: {Self-Guide}. Please verify your reasoning process for errors based on this analysis, then refine your reasoning process and answer.
Generation Format: 1. Inference process: 2. Answer:

**得到的最终答案**
**Inference process:**
**1.** Identify the family members mentioned in the story.
**2.** Determine the relationships between the family members that are explicitly mentioned in the story.
**3.** Identify any indirect relationships that may help determine the relationship between the two family members in question.
**4.** Consider the relationships between the family members and how they might connect the two family members in question.
**5.** Verify the reasoning process for errors based on the analysis and refine if necessary.
**Answer:** Based on the analysis, Dorothy is the grandmother of Oliver.

Table 8: 案例分析。我们随机选择了一个CLUTRR数据集中的例子来展示Self-Guide的效果，并突出显示了答案和相关内容。这是英文原始输入和输出内容。


(c) 2024 中国中文信息学会计算语言学专业委员会

**问题：Jane was born on the last day of February in 2000. Today is her 16-year-old birthday. What is the date yesterday in MM/DD/YYYY?"**
**答案：**02/28/2016

**原始推理链（CoT）输入提示**
To solve the problem, Please think and reason step by step, then answer. Question: {Question} Please format the date in the following manner: MM/DD/YYYY.
Generation Format: 1. Reasoning process: 2. Answer:

**原始推理链（CoT）输出回答**
Reasoning process:
**1.** Jane was born on February 29, 2000, which was a leap year.
**2.** Since today is her 16th birthday, we need to calculate the date 16 years ago.
**3.** 16 years ago from 2021 is 2005.
**4.** Since 2005 was not a leap year, February 29th did not exist. Therefore, her birthday would have been on February 28th.
Answer: 02/28/2005

**步骤1：基于Self-Guide的自我规划**

**获取Self-Guide的输入提示**
You are a knowledgeable and patient professor whose role is to guide students in solving problems correctly. Here is a question:{Question} please provide a detailed analysis. Note: Since your responsibility is to guide students in answering the question, your analysis should think step by step, Please note that your role is to guide them step by step through the problem, so please don't give them the final result.

**得到的Guideline**
**Step1:** Determine the current date. Today is Jane's 16th birthday, so we need to calculate the current date. Since she was born on the last day of February in 2000, her birthday would be on February 29th, 2016 (since 2000 was a leap year).
**Step2:** Calculate yesterday's date.To find yesterday's date, we need to subtract 1 day from Jane's birthday. Since February 2016 has 29 days, we need to subtract 1 from 29 to get yesterday's date.
**Step3:** Consider the month and year after subtracting 1 from 29.
**Step4:** Finalize the answer now that we have calculated yesterday's date. By following these steps, we can determine the date yesterday in MM/DD/YYYY format.

**步骤2：最终答案生成**

**获取最终答案的输入提示**
{CoT-log}
This is a credible analysis of this question: {Self-Guide}. Please verify your reasoning process for errors based on this analysis, then refine your reasoning process and answer.
Generation Format: 1. Inference process: 2. Answer: MM/DD/YYYY.

**得到的最终答案**
**Inference process:**
**1.** The current date is February 29, 2016, as it is Jane's 16th birthday and she was born on the last day of February in 2000, which was a leap year.
**2.** To find yesterday's date, we subtract 1 day from February 29, 2016, which gives us February 28, 2016.
**3.** Considering the month and year, yesterday's date is February 28, 2016.
**Answer:** 02/28/2016

Table 9: 案例分析。我们随机选择了一个Date数据集中的例子来展示Self-Guide的效果，并突出显示了答案和相关内容。这是英文原始输入和输出内容。


(c) 2024 中国中文信息学会计算语言学专业委员会

# 基于大模型的交互式谎言识别：数据和模型

纪程炜[1,†]，王思远[1,†]，李太山[1,†]，牟馨忆[1]，赵丽敏[1]，
薛兰青[2]，应缜哲[2]，王维强[2]，黄萱菁[3]，魏忠钰[1,*]
[1]复旦大学　大数据学院，上海市　200433
[2]蚂蚁集团，杭州市　310023
[3]复旦大学　计算机科学技术学院，上海市　200433
21210980108@m.fudan.edu.cn

## 摘要

面向对话交互过程的谎言识别技术在不同的应用场景有广泛的应用需求。现有的鉴谎技术往往在整体的对话级别上给出最终决策，而缺乏对细粒度谎言特征和线索的逻辑分析，难以满足场景中对于可解释性的需求。本文提出了谎言指征和语义不一致线索的概念，用于帮助识别对话中的谎言，提升鉴谎方法的可解释性。文章同时提出一个谎言识别框架，用于训练谎言识别大语言模型（LD-LLM）。它利用细粒度的谎言指征并且发现对话中是否存在语义不一致线索，以实现更可靠的谎言识别。文章在真实交互场景中构建了两个谎言识别数据集**FinLIE**和**IDLIE**，分别关注金融风控场景和身份识别场景。实验结果表明，基于这两个数据集创建的指令数据集微调得到的LD-LLM，在基于真实交互的谎言识别上达到了最先进的水平。

关键词：　谎言指征；语义不一致线索；**FinLIE**；**IDLIE**；LD-LLM

# Unveiling Lies: Enhancing Large Language Models for Real-World Lie Detection in Interactive Dialogues

**Chengwei Ji[1,†], Siyuan Wang[1,†], Taishan Li[1,†], Xinyi Mou[1], Limin Zhao[1], Lanqing Xue[2], Zhenzhe Ying[2], Weiqiang Wang[2], Xuanjing Huang[3], Zhongyu Wei[1,*]**
[1]School of Data Science, Fudan University, Shanghai 200433
[2]Ant Group, Hangzhou 310023
[3]School of Computer Science, Fudan University, Shanghai 200433
21210980108@m.fudan.edu.cn

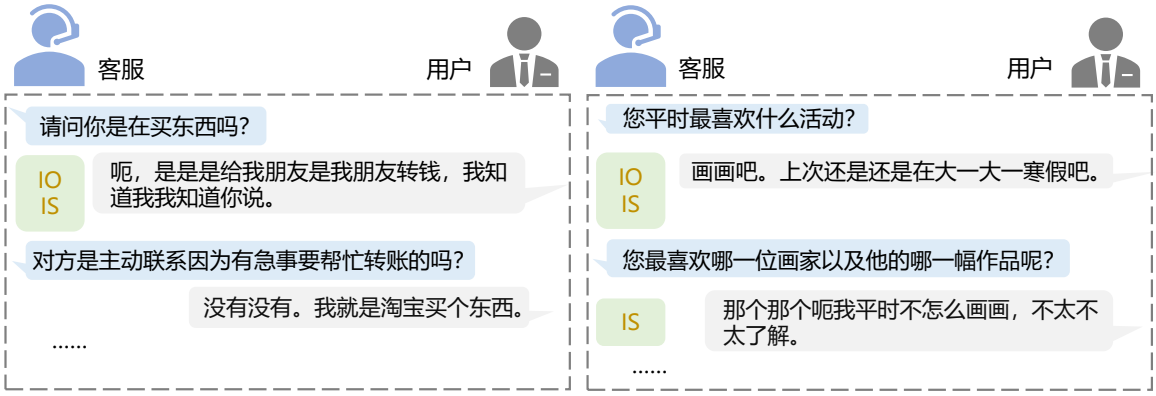
## Abstract

Lies are so common in everyday conversation, and existing techniques for identifying lies from the dialogue give a final decision without explanations. They lack logical analysis of fine-grained deceptive features for dialogues. In this paper, we propose the concept of deception indicators and semantic inconsistency clues for detecting lies. We introduce a lie detection framework to train a Lie Detection LLM (LD-LLM), which utilizes deception indicators and elicits inconsistency clues for more reliable lie detection. We construct two lie detection datasets **FinLIE** and **IDLIE** in real-world scenarios focusing on financial risk control and identity detection. The LD-LLM is then fine-tuned using an instruction-tuning set created based on these two datasets. Experimental results indicate that our framework achieved SOTA.

**Keywords:** Deception indicators , Clues of semantic inconsistency , FinLIE , IDLIE , LD-LLM.

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

## 1 引言




(a) FinLIE (b) IDLIE

Figure 1: 来自**FinLIE**和**IDLIE**数据集的两个谎言对话示例。在第一种场景下，用户回答问题时提供了过多的信息，并在说话断断续续。此外，用户首先声称将钱转给朋友，后来又改成购买产品，显示出他陈述中的明显不一致。在第二种情况下，用户在回答问题时不仅说话不连贯、提供了无用信息，而且前后内容出现了不一致现象。

面向对话交互过程的谎言识别技术在不同的应用场景有广泛的应用需求。例如，金融公司的支付平台的智能客服可能会询问用户有关交易细节从而判断用户的交易是否存在风险，而用户会通过说谎来掩盖他们的真实交易目的，以谋取个人利益。此外，违法者经常利用虚假身份从事欺诈活动，以获取不合法的财富。这些欺骗行为将导致财产损失，并对公共安全构成威胁。此外，最近大语言模型（LLMs (Touvron et al., 2023; Ouyang et al., 2022; Touvron et al., 2023)）通过和人类的交互对话来解决众多问题，但不安全和不可靠的模型可能会存在欺骗人类的风险。基于此，在交互对话中检测谎言具有重要意义。

基于交互对话的谎言识别主要存在如下三个挑战：（1）现实中交互式对话的谎言识别技术在不同的应用场景有广泛的应用需求，如金融交易平台需要通过客服与用户的对话判断用户是否说谎，以保障平台用户的财产安全。如何利用模型来帮助解决现实世界中的谎言识别问题是极具挑战的。（2）缺乏反映真实场景的对话交互谎言识别数据集。如表1所示，虽然一些现有数据集(Gros et al., 2021a)包含多轮欺骗性对话，但它们是为了检测与人类进行对话的是否是智能机器人而设计的。这与真实场景中人类的欺骗行为不一致。（3）之前的谎言识别不具备动态特性，与现实场景不相符。现有的鉴谎技术(Wang, 2017; Conroy et al., 2015; Vaccari and Chadwick, 2020; Peskov et al., 2020)往往在整体的对话级别上给出最终决策，而缺乏对细粒度谎言特征和线索的逻辑分析，从而缺乏可解释性。本文提出了谎言指征和语义不一致线索的概念，用于帮助识别对话中的谎言，提升鉴谎方法的可解释性。

为了提升真实对话场景中的自动化谎言识别能力，本文构建了两个数据集，**FinLIE**数据集和**IDLIE**数据集。**FinLIE**数据集专注于金融风控领域，旨在确定用户是否在各种金融交易细节上说谎。**IDLIE**数据集专门用于检测用户是否在身份信息上说谎。图1展示了两个数据集中欺骗性样例，对话流中存在着语义不一致现象，比如图1 (a)中用户先回答给朋友转账，后面改成了网购，而图1 (b)中用户先回答喜欢绘画，后面回答“平时不怎么画画”，存在明显的前后不一致；同时对话流中存在细粒度表达特征，比如提供过多信息和犹豫不决。这些特征共同表明用户很可能在进行欺骗行为。

为了建模对话中的谎言，本文提出基于大语言模型的对话交互谎言识别方法。近期研究(Brown et al., 2020; Chowdhery et al., 2023)表明，大语言模型在多种下游任务中展现出了显著的性能，特别是在交互式对话情境下，它们能够有效地捕捉历史信息和理解语义。尽管如此，利用大语言模型在对话中进行谎言识别仍具挑战性。为了提高这一能力，本文基于提出的**FinLIE**和**IDLIE**数据集构建了一个指令微调数据集，并使用该指令微调集对大语言模型进行微调。此外，采取两项措施来精确建模对话中的谎言细粒度特征：首先，为了增强对对话语义一致性方面的分析能力，在对话谎言识别指令微调数据集的输出中加入了语义不一致的逻辑线索。具体而言，采用思维链策略(Wei et al., 2022)，通过提示ChatGPT 来获取每段欺骗性对话中


(c) 2024 中国中文信息学会计算语言学专业委员会

的关键逻辑不一致点。其次，基于谎言识别理论，引入了谎言指征的概念，通过结合语义线索之外的细粒度特征，例如说话风格和语气助词等，来用于分析谎言。我们自动识别对话中每个子句的谎言指征，并将它们作为指令微调数据集的额外输入。通过这两种措施，本文扩展了对话谎言识别的指令微调数据集，使其包括交互对话中的谎言指征和语义不一致线索，以此来微调大语言模型，从而实现更可靠的谎言识别。因此本文的主要贡献有两方面：

- 本文是首次提出在交互过程中识别谎言，同时提出谎言指征和对话中语义不一致线索的概念，将其结合来帮助谎言检测。与传统的谎言识别的方法不同，我们的方法通过在细粒度上分析对话内容的一致性，从而更有效地揭示谎言，也提高了鉴谎的可解释性。
- 本文采用了微调大语言模型的方法，该方法是目前自然语言处理领域的前沿技术之一。同时，我们构建了两个新的针对真实场景（金融风控和身份识别）的谎言识别数据集，**FinLIE**和**IDLIE**。这类针对真实场景的交互式对话的谎言数据集在谎言识别研究中非常稀少。不同于其他指令微调数据，针对谎言识别，我们在数据中引入了谎言指征和对话中的语义不一致线索，这使得我们的模型能够捕捉到谎言行为中的细微语言特征和语义不一致，这是其他模型所不具备的。通过和多个基线模型的对比，我们的模型能够显著优于现有的通用大模型，同时也能取得微调其他轻量模型不能达到的效果。

## 2 数据集

### 2.1 数据集构造

由于现有研究中缺乏在真实场景下的多轮对话数据集，本文构建了两个数据集：**FinLIE**和**IDLIE**，旨在研究在金融风控和身份识别领域中的多轮对话谎言识别。对比以往数据集，如表1所示，先前的数据集只涉及真实场景或多轮对话，例如**R-U-A-Robot**(Gros et al., 2021b)以及**Diplomacy**(Peskov and Cheng, 2020)数据集符合多轮对话，但是对话不是基于真实场景的。**Deceptive Opinion spam Corpus v1.4** 以及**DECOUR**(Fornaciari and Poesio, 2013)数据集符合真实场景，但是均为单轮对话。相比之下，**FinLIE**以及**IDLIE**数据集具备真实场景下的多轮对话特点，因此具有明显优势。

| 数据集 | **R-U-A-Robot** | **Deceptive Opinion Spam Corpus v1.4** | **Diplomacy** | **DECOUR** | **FinLIE** | **IDLIE** |
|---|---|---|---|---|---|---|
| 是否真实场景 | | ✓ | | ✓ | ✓ | ✓ |
| 是否多轮对话 | ✓ | | ✓ | | ✓ | ✓ |

Table 1: 不同数据集与**FinLIE**和**IDLIE**的对比。

在一家互联网金融公司的支持下，本文构建了金融场景下的**FinLIE**数据集，包括了客服与用户之间关于交易风险调查的对话。经过与金融公司的沟通，我们将会提供脱敏版本 。具体地，当用户在进行大额转账时会触发风控机制，智能客服会询问用户具体的交易细节以及交易对象，该客服会提醒用户交易过程中可能存在风险，以减少用户的损失。为了简化对话中的谎言识别过程，文章基于语言学和欺诈检测理论(羊芙葳, 2010)，提出了子句级别的谎言指征作为辅助信息，并对谎言做如下定义：说谎是人们有意把自认为虚假的信息传递给他人的言语行为与交际策略，谎言是说谎的产物。文章定义了六种子句级别的谎言指征：直接挂断或中途挂断（HDH）、说话不连贯（IS）、拒绝回答（RA）、信息过量（IO）、答非所问（IA）、喜用语气词（LUMP）。每个子句可能符合一个或多个谎言指征，对于每段完整对话，在子句级别标注了谎言指征以及在对话级别上标注了谎言标签。详细的标注过程包括三个步骤。首先，将用户和客服之间的每一轮对话视为一个子句，并相应地将对话分割成多个子句。其次，每段对话及其子句分配给两名专业的标注员，以获取谎言指征和谎言标签。如果两名标注员意见不一致，将按照多数原则由第三名标注员进行检查。如果任何两名标注员无法达成一致意见，则丢弃该实例。最后，本文计算两名标注员之间的Kappa系数以评估标注质量，并计算了平均谎言识别的Kappa系数为**74.7%**，谎言指征预测的Kappa系数为**79.1%**，表明结果均为可靠的。

https://myleott.com/op-spam.html
https://github.com/jcwzs/Lie_Detection


(c) 2024 中国中文信息学会计算语言学专业委员会

此外，本文创建了一个基于日常对话中的身份识别场景，并构建了**IDLIE**数据集。首先，设计一个包含四个类别的身份库：体育爱好者、社区成员、艺术爱好者、其他，每个类别都包含多个子身份。例如，篮球迷、足球迷和羽毛球迷都属于体育爱好者。接着，建立了客服与用户之间的对话，旨在检测用户的身份，并进行谎言识别。每个用户一定属于身份库中预定义的一个身份，来作为他们的真实身份。随后，他们选择另一个从未拥有过的身份作为虚假身份。客服会针对真实和虚假身份提问。客服不知道用户的身份是真实还是虚假的，会选择几个问题来询问用户以判断身份的真实性。对于每一次对话，客服专注对话级别的谎言识别，以确定用户的身份是真实还是虚假的。用户只会在其身份是虚假的情况下说谎，即，当客服询问与真实身份相关的问题时，用户不会采取欺骗行为；当针对虚假身份提问时，用户会说谎。在**IDLIE**数据集中，为了节约成本，本文利用在**FinLIE**数据集上经过充分训练的轻量模型预测谎言指征。由于特殊的设定，HDH和RA在**IDLIE**中不存在，因此其预测结果为零。在对话结束之后，客服需要判断对话过程中用户是否说谎。整个实验过程招募了40名本科生，分别扮演用户和客服的角色。为了确保实验的可靠性，扮演客服角色的学生可以参与多次实验，而扮演用户角色的学生只能参与不超过两次实验。采取这种设置是为了与实际情况保持一致，客服通常具有丰富的提问经验。而用户因为只参与少量实验，并不具备高超的撒谎能力。整个实验过程中在线会议中进行，并限制最大对话轮数为10次。在线会议的语音信息通过自动语音识别技术（ASR(Li and others, 2022)）转换为文本。最终**IDLIE**数据集包含720段对话和4,001个子句。

## 2.2 数据集分析

表2中展示了两个数据集的基本统计信息，包括两个数据集中谎言和非谎言样例的分布情况。在**FinLIE**数据集中，有498段对话包含谎言，占比59.9%。在**IDLIE**数据集中，有360段谎言对话，占比50.0%。谎言对话的平均长度比非谎言要长，表明用户在说谎过程中可能更倾向于提供更多的额外信息。

| 是否说谎 | **FinLIE** | | **IDLIE** | |
|---|---|---|---|---|
| | 数量 | 平均长度 | 数量 | 平均长度 |
| 谎言 | 498 | 281.3 | 360 | 486.2 |
| 非谎言 | 334 | 225.2 | 360 | 481.1 |

Table 2: 两个数据集的基本统计信息。平均长度表示对话的平均字数。

表3 (a)中展现了**FinLIE**数据集不同谎言指征的统计数据。每个指征的比例相差较大，这使得谎言指征预测任务具有挑战性。在各类指征中，HDH占比最大，因为一些用户可能认为与客服沟通是浪费时间。尽管LUMP的比例最少，但它可以显示用户在隐瞒一些事情时表现紧张，使得LUMP成为判断用户是否在说谎的重要标志。IO意味着用户会提供一些无用的信息，因此平均长度最长。HDH和RA意味着用户不愿意回答客服的问题，通常会给出相对较短的回复。表3 (b)列出了**IDLIE**数据集预测后的统计信息。由于HDH和RA两个指征在**IDLIE**构建过程中不存在，其预测结果为零。我们从**IDLIE**数据集中抽取了30段对话进行人工标注，与Re-BERT预测结果和BERT的预测结果分别计算F1值，结果如表4所示，自动化标注的F1分数在**70%**以上，可以作为下游自动化标注的工具。

| 统计信息 | **HDH** | **IS** | **RA** | **IO** | **IA** | **LUMP** |
|---|---|---|---|---|---|---|
| 子句数量 | 382 | 317 | 305 | 202 | 163 | 29 |
| 比例 | 9.0 | 7.5 | 7.2 | 4.8 | 3.9 | 0.7 |
| 平均长度 | 30.7 | 52.9 | 36.1 | 61.6 | 43.9 | 57.9 |

(a) **FinLIE**

| 统计信息 | **HDH** | **IS** | **RA** | **IO** | **IA** | **LUMP** |
|---|---|---|---|---|---|---|
| 子句数量 | - | 280 | - | 136 | 140 | 36 |
| 比例 | - | 7.0 | - | 3.4 | 3.5 | 0.9 |
| 平均长度 | - | 87.9 | - | 82.3 | 83.6 | 81.7 |

(b) **IDLIE**

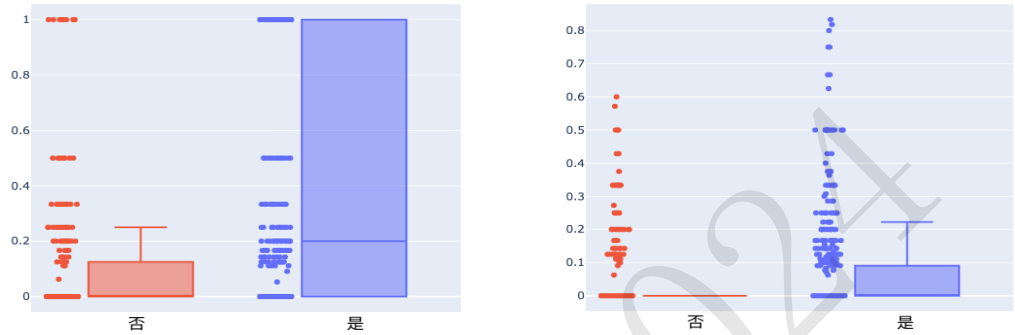
Table 3: 两个数据集的谎言指征统计信息。平均长度表示每条子句的字符数，比例表示谎言指征占总子句数量的百分比。


(c) 2024 中国中文信息学会计算语言学专业委员会

| 模型 | 平均F1 | HDH | IS | RA | IO | IA | LUMP |
|---|---|---|---|---|---|---|---|
| Re-BERT | **73.0** | 87.5 | 75 | 82.9 | 76.7 | 75.8 | 40.2 |
| BERT | 70.8 | 80.0 | 72.9 | 90.0 | 61.1 | 75.3 | 45.2 |

Table 4: 人工检测对比结果。

除了基本统计信息，本文进一步探讨了**FinLIE**数据集中谎言指征和谎言标签之间的关系。图2（a）和（b）展示了谎言指征（HDH和RA）与是否说谎之间的箱线图，这两个指征与谎言标签之间的关系最为显著。横轴表示用户是否说谎，纵轴表示谎言指征的频率。图2（a）显示了HDH与说谎之间的相关性，表明HDH的频率越高，用户说谎的可能性就越大。图2（b）显示了RA与说谎之间的相关性，RA的频率与说谎的可能性呈正相关。这两个谎言指征都表明，不愿意回答客服问题是说谎的重要标志。




(a) 直接挂断或中途挂断　　(b) 拒绝回答

Figure 2: 不同谎言指征与是否说谎的相关性分析。

## 3 方法

### 3.1 总览

谎言识别框架如图3所示。在给定一段对话后，预测对话中可能存在的谎言指征，并将这些指征嵌入到对话中作为输入。通过微调谎言识别大模型，分析对话是否存在语义不一致线索，再基于此确定用户是否说谎。本节先介绍谎言指征预测，见第3.2节，接着介绍指令数据集的构建过程，见第3.3节，最后描述谎言识别大语言模型的微调过程，见第3.4节。

### 3.2 谎言指征预测

为了识别六种谎言指征，本文提出一种轻量级的Recurrent BERT模型(Hong et al., 2021)，进行多标签分类的谎言指征预测。该模型能有效地利用上游对话信息，在每个单轮对话中预测谎言指征。具体而言，该模型将原始对话分为单独的子句，将这些子句按照时间顺序视为序列输入，输出则是对应于每个子句的谎言指征标签。此外，该模型不仅能够预测用户在对话每一轮中的谎言指征，还能够利用全局上下文信息，在对话级别上判断用户是否在说谎，以补充谎言指征的预测。谎言指征预测模型使用**FinLIE**数据集进行训练，并预测**IDLIE**数据集中的谎言指征。为了确保充分训练，Focal Loss (Lin et al., 2017)被选择作为预测谎言指征的损失函数。

### 3.3 指令数据集构造

为了让大模型能够分析对话中的语义不一致线索，并判断用户是否在说谎，本文构建一个谎言识别指令微调数据集，并基于此微调大语言模型。指令数据集中的样本以"输入-输出"对的形式存在。输入包含原始对话和谎言指征，以及要求模型识别对话中谎言的指令。输出包括一段分析的语义不一致线索和基于此预测的谎言标签，表明用户是否在说谎。所有多轮对话具备统一格式的指令，具体地，每个子句之后都加入了相应谎言指征的文本描述，以此构成完整的


(c) 2024 中国中文信息学会计算语言学专业委员会

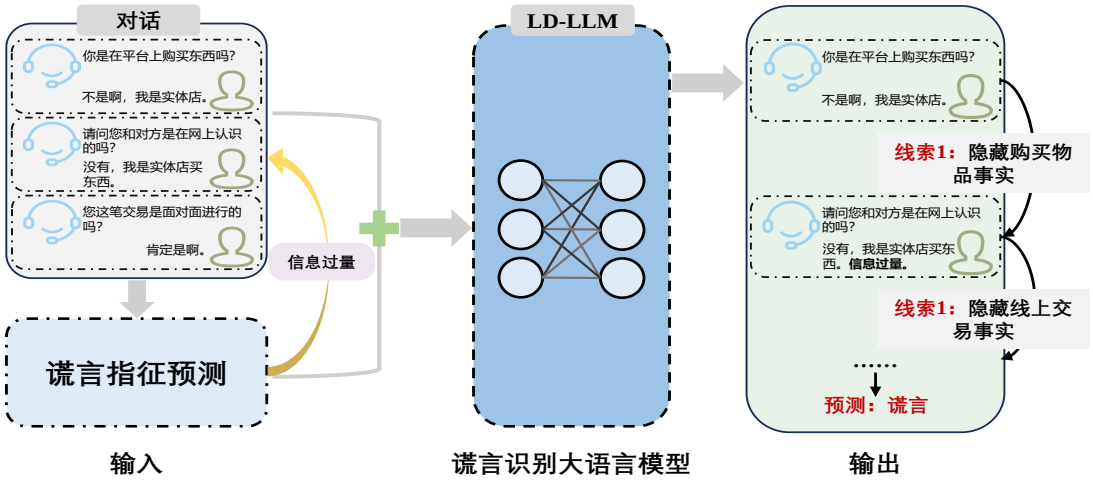



Figure 3: 谎言识别框架图。谎言识别框架图展示了谎言指征预测模块。对于输入的一段对话，该模块为对话中的每个子句预测谎言指征，并将其作为附加输入。谎言识别大语言模型（LD-LLM）分析对话内的语义不一致的线索，并利用大语言模型的自回归属性根据输出的语义不一致线索来判断用户是否说谎。

| | | |
|---|---|---|
| 输入 | 指令 | 欺骗（英语：Deception）是一种误导他人、隐瞒真相，或者推广假的想法、概念或信念的一种行为或陈述。你的任务是分析客服人员与用户之间的对话，以在对话中搜寻表明用户在欺骗的线索，最终确定用户是否在欺骗。 |
| | 对话 | 客服："请问你是在买东西吗？"用户："呃，是是是我朋友是我朋友，我知道我我知道你说，给我朋友转账。"，说话不连贯、信息过量……<br>客服："请问您是在购买什么商品呢？"用户："不好说。"，拒绝回答……<br>客服："请问您和对方是在网上认识的吗？"用户："呃，不是在网上，我不不认识他"，说话不连贯<br>客服："那您是因为什么原因付款呢？"用户："嗯，帮他买的东西，我们是朋友吧。"…… |
| 输出 | 分析 | 在这段对话中，有几个线索表明用户在欺骗。<br>首先，当客服问到用户购买什么商品时，用户拒绝回答。这可能是因为用户不希望透露真实的交易目的。……其次当客服问到用户是否在网上认识对方的时候，用户回答不认识对方，但是在对话开始就已经回答给朋友转账，**与之前回答有着明显的不一致**。<br>最后，当客服问到用户付款原因时，用户的回答是"嗯，帮他买的东西，我们是朋友吧。"说明用户在给朋友付款，而之前又回答不认识对方，**再次出现明显的不一致现象**，表明用户可能在隐藏交易对象，掩盖用户真实的行为。<br>用户多次前后不一致、拒绝回复客服问题等行为都是常见的欺骗手段。……这些线索表明用户在对话中存在欺骗行为。 |
| | 标签 | 1 |

Figure 4: **FinLIE**数据集中的指令微调实例。拒绝回答表示用户没有回答客服的问题。信息过量表示用户除了回答客服的问题，还说了一些无关的内容。说话不连贯表示用户说话断断续续，没有逻辑性。在实例中，指令和对话（谎言指征嵌入在原始对话内）作为模型的输入，而对于对话的逻辑分析和基于逻辑分析预测的标签则做为模型的输出。

输入。模型的输出包括针对多轮对话的逻辑分析，以及根据生成的分析内容判断用户是否在说谎，指令数据集中的一个具体样例如图4所示。在多轮对话中，谎言通常表现为对话不同部分之间的语义不一致。为了捕捉这样的特性，本文采用思维链策略，利用ChatGPT自身的推理能力，自动提取这些谎言中的不一致线索。具体的，向ChatGPT提供了一段对话，并告知其中存在说谎行为以及对话中的谎言指征。然后，令其分析对话，并识别谎言中的任何不一致之处。为了防止过多的干扰，将分析内容的长度限制在700个汉字以内。具体提示词如下：

语义不一致线索生成的提示

**system:** "你是李华，一个欺骗检测的专家。你的任务是分析客服人员与用户之间的对话，并在对话中找到表明用户在欺骗的文字片段。"
**user:** "欺骗（英语：Deception）是一种误导他人、隐瞒真相，或者推广假的想法、概念或信念的一种行为或陈述。我会给你一段对话，在这段对话中，我们已经确认用户在欺骗。你需要一步一步地分析，从对话原文中找到用户说谎的线索，并给出解释。明白了吗？"
**assistant:** "好的，我明白了。我是李华，我会分析你给的对话。"
**user:** "很好，那我们开始吧。对于给定的对话：
{conversation}
对话中是否有任何线索表明用户在欺骗？请给出线索和理由。你的答案必须在700字以内。答案：让我们一步一步地思考，"


(c) 2024 中国中文信息学会计算语言学专业委员会

### 3.4 指令微调

LLaMa-2被用来作为基座，使用基于**FinLIE**和**IDLIE**数据集构造的指令数据集，结合LoRA(Hu et al., 2021)技术对LLaMa-2模型进行微调。LoRA的原理是在模型的原始权重上添加了低秩调整，我们将LoRA应用于自注意力模块中的所有查询/键/值/输出投影矩阵。

## 4 实验

### 4.1 实验设置

基于**FinLIE**数据集和**IDLIE**数据集构建指令微调数据集，分别包含832个实例和720个实例，随后利用指令微调数据集对LLMs进行微调。数据集按8:1:1的比例被划分成了训练集、验证集和测试集。7B版本的Chinese-LLaMA-Alpaca-2 做为基础模型，并且利用AdamW(Loshchilov and Hutter, 2017; Loshchilov and Hutter, 2018)优化器进行优化。训练20轮，数据批量大小为3，学习率为1e-4，最大输入长度为900。

### 4.2 基线模型

LD-LLM与以下基线模型进行了比较。首先，本文训练了四个以BERT(Devlin et al., 2018)为基础的轻量模型进行对比：

1. *BERT*：仅仅将原始对话做为BERT的输入，输出对话中是否包含谎言。
2. *BERT+LI*：将谎言指征与原始对话进行拼接做为BERT的输入。
3. *Diff-W*：将原始对话以单轮对话的形式输入到BERT，为每一单轮对话学习权重，并且以加权的方式表示。
4. *Re-BERT*：谎言可能在整个对话过程中发生，不同回答之间存在语义或者内容上的不一致性。为了充分利用先前的信息并模拟其与当前子句的交互，文章实施了一种基于Transformer的循环分类模型，称为Recurrent BERT模型（Re-BERT）。

同时还比较了当前的几个大语言模型，包括：

1. *LLaMa*：直接利用提示学习让LLaMa-2进行推理，无需训练。在推理过程中，采用了2-shot学习，向模型提供一个正样本和一个负样本作为样例。
2. *ChatGPT*：结合思维链，向ChatGPT提供几个样例，提示其进行推理。
3. *LLaMa w/ fine-tuning*：使用原始指令微调数据集对LLaMa-2进行微调，通过提供一个指令，该数据集仅使用原始对话以及谎言标签作为模型的输入和输出。

在使用ChatGPT进行谎言识别时，采用了思维链策略，并结合few-shot提示来充分发挥其推理能力。具体来说，在提示样例中，提供了样例的谎言指征以及相应的谎言线索。提示的详细信息如下。从实验结果中可以看到，ChatGPT的性能不如LD-LLM模型甚至是不如轻量的Re-BERT模型。通过进一步地分析发现，在对说谎的例子进行预测时，ChatGPT时常会出现类似“没有明显的线索”、“没有出现明显的矛盾或模糊之处”的输出。”这表明在处理谎言识别这类复杂任务的时候，未经微调的大模型并不能很好地捕捉到用户在说谎的线索及指征，因而不能识别出谎言，说明对于谎言识别去构建领域指令数据来微调模型的必要性。

**利用ChatGPT进行谎言识别的提示。**

你是李华，一个欺骗检测的专家。你的任务是分析客服人员与用户之间的对话，以判断对话中的用户是否在说谎。请你严格按照如下格式来回复，你的回复必须以“因此，答案为：是/否”为结尾。
对话：<样例1>
是否说谎：<样例1的思维链及答案>
对话：<样例2>
是否说谎：<样例2的思维链及答案>
对话：
{conversation}
是否说谎：

https://github.com/ymcui/Chinese-LLaMA-Alpaca-2


(c) 2024 中国中文信息学会计算语言学专业委员会

| 模型 | | FinLIE | | IDLIE | |
|---|---|---|---|---|---|
| | | 宏观F1 | 准确率 | 宏观F1 | 准确率 |
| 轻量模型 | BERT | 54.8 | 60.4 | 41.0 | 47.2 |
| | BERT+LI | 76.0 | 71.4 | 44.7 | 52.5 |
| | Diff-W | 84.3 | 80.0 | 62.1 | 63.7 |
| | Re-BERT | 85.2 | 81.3 | 72.1 | 72.2 |
| 未训练的大模型 | LLaMa | 47.3 | 59.0 | 35.7 | 48.6 |
| | ChatGPT | 64.9 | 68.1 | 35.7 | 50.0 |
| 微调大模型 | LLaMa w/ fine-tuning | 83.1 | 81.7 | 73.5 | 73.6 |
| | **LD-LLM** | **86.1** | **85.6** | **75.3** | **75.6** |
| | **Human Performance** | - | - | **76.9** | **77.0** |

Table 5: 段落级别的谎言识别任务上的所有实验结果。实验结果均包含**FinLIE**数据集以及**IDLIE**数据集上的宏观F1得分以及准确率。所有的评价指标均为百分比格式。其中Human Performance仅对于**IDLIE**数据集进行实验。

此外，作者对数据集的一部分进行了标注，构成了**IDLIE**数据集上的人类表现（Human Performance），用来确定一个上限。具体而言，随机选取了100段对话，这其中包含了测试集的全部数据，并对其进行人工标注，**F1**得分为**76.9**。

### 4.3 谎言识别结果总览

表5展示了以BERT为主体模型及其各种变体以及各种大语言模型在**FinLIE**数据集和**IDLIE**数据集上进行谎言识别的结果。实验结果表明，本文提出的谎言指征和语义不一致线索相较于其他基线模型，显著改善了实验结果，验证了框架的有效性。这是因为谎言指征可以在句子级别上提供更加细粒度的特征，使得大语言模型能够在子句级别上捕捉到相关信息。另一方面，语义不一致的线索为每段欺骗性对话提供了合理的解释，使得模型可以更好地理解对话，利用大语言模型的自回归属性，使得模型能够根据生成的线索判断用户是否在说谎，从而提高大语言模型在识别谎言的准确性和F1得分。同时，文章在身份识别领域下比较了人类表现与LD-LLM。如表5所示，表明在谎言识别任务上，人类表现要优于LD-LLM，相比于人类在这一类复杂任务上的优秀表现，大模型还有提升空间。

### 4.4 消融实验

| 模型 | FinLIE | | IDLIE | |
|---|---|---|---|---|
| | 宏观**F1** | 准确率 | 宏观**F1** | 准确率 |
| LLaMa w/ fine-tuning | 83.1 | 81.7 | 73.5 | 73.6 |
| LD-LLM w/o I | 83.9 | 83.0 | 74.9 | 75.0 |
| LD-LLM w/o C | 84.7 | 83.2 | 73.0 | 73.1 |
| **LD-LLM** | **86.1** | **85.6** | **75.3** | **75.6** |

Table 6: 在**FinLIE**数据集与**IDLIE**数据集上的消融实验分析。

本小节通过表6探索了不同模块的作用。LD-LLM w/o I 表示模型在输出中包括对于对话的逻辑分析，输入中不包括谎言指征。LD-LLM w/o C 表示输入中包括谎言指征而模型并不输出对于对话的逻辑分析。实验结果表明，为欺骗性对话生成语义不一致线索以及谎言指征的方法可以显著提高大语言模型的谎言识别准确率和F1得分。从对话中提取子句级别的细粒度特征，并进行逻辑分析以确定用户是否在说谎是非常有效的。


(c) 2024 中国中文信息学会计算语言学专业委员会

| 模型 | BERT | Diff-W | **Re-BERT** |
|---|---|---|---|
| 平均**F1** | 75.4 | 71.8 | **75.9** |
| **HDH** | **99.3** | 97.2 | 91.2 |
| **IS** | 70.7 | 66.8 | **78.1** |
| **RA** | **92.5** | 84.8 | 91.4 |
| **IO** | 67.2 | 62.6 | **70.4** |
| **IA** | 73.1 | 66.9 | **74.2** |
| **LUMP** | 49.6 | 49.7 | **49.9** |

Table 7: 在**FinLIE**数据集上的谎言指征预测任务的实验结果。

### 4.5 谎言指征预测

为了展示谎言指征预测有效性，表7展示了使用不同轻量模型在**FinLIE**数据集上进行谎言指征预测的相应结果。实验结果表明，Re-BERT模型在整体上可以达到比较高的谎言指征预测准确率，这表明提供过多的后续信息可能会降低预测谎言指征的有效性，同时Re-BERT模型更符合现实对话场景，说明该模型在**IDLIE**数据集上对谎言指征的预测展现了很好的泛化能力，并在谎言识别任务上表现良好如表5所示，证明了整体框架的可靠性。

### 4.6 泛化分析

| 模型 | **IDLIE**数据集 | | 酒店数据集 | |
|---|---|---|---|---|
| | 宏观**F1** | 准确率 | 宏观**F1** | 准确率 |
| GPT-3.5 | 35.7 | 50.0 | 48.3 | 57.0 |
| GPT-4.0 | 37.7 | 50.0 | 45.5 | 53.0 |
| LLama | 52.7 | 51.8 | 52.7 | 51.8 |
| Re-BERT | 52.6 | 50.1 | 52.6 | 50.1 |
| **LD-LLM** | **66.4** | **68.8** | **67.0** | **69.3** |

Table 8: 在两个数据集上LD-LLM与其他基线模型的性能对比结果。

为了验证框架的通用性，文章在两种不同的泛化设定下进行了实验。首先，在IDLIE数据集上，对于陌生身份进行了泛化实验。数据集中，一些特殊身份只存在于真实身份（例如台球迷）或者只存在于虚假身份（例如游戏迷）的样本数量不超过三个。因此，在身份识别场景中，采用了启发式规则进行数据集拆分。具体来说，文章统计了数据集中出现五次或更少的身份，例如环境工作者，并从中挑选出这些身份所对应的样本，形成了新的测试集，最终测试集包括了63个样本。表8展现了基线模型和LD-LLM在谎言识别方面的对比结果。尽管模型性能普遍下降，但LD-LLM优于基线模型，表明谎言指征和语义不一致的线索能够捕捉与谎言相关的标志，而不仅仅是与身份相关的信息。

接着为了检验框架具有跨领域的泛化能力，我们将在**FinLIE**数据集上训练得到的LD-LLM应用到其他领域上进行实验。由于公开对话谎言数据的稀缺，我们选择一个欺骗性评论的数据集(Ott et al., 2011)进行实验。该数据集包含了针对一些酒店的真实和虚假评论，我们将其转换为单轮对话的格式，并在经过训练的LD-LLM和基线模型上进行实验。如表8所示，在酒店数据上，LD-LLM的性能显著地优于基线模型。所以我们的LD-LLM相较于其他模型在陌生身份和跨领域设定上都有较好的泛化性。

### 4.7 样例分析

图5展示了LD-LLM的分类正确的样例。可以观察到，模型生成的语义不一致分析接近于期望输出，并且最终能够合理地判断用户在说谎。这个例子同时也是LLaMa w/ fine-tuning分类错误的样例，突出了在句子级别结合细粒度特征以及利用模型生成的逻辑分析对对话进行增强，可以有效提升模型的识别谎言的性能。


(c) 2024 中国中文信息学会计算语言学专业委员会

| | | |
|---|---|---|
| 输入 | 指令 | 欺骗（英语：Deception）是一种误导他人、隐瞒真相，或者推广假的想法、概念或信念的一种行为或陈述。你的任务是分析客服人员与用户之间的对话，以在对话中搜寻表明用户在欺骗的线索，最终确定用户是否在欺骗。 |
| | 对话 | 客服：“请问你是在买东西吗？”用户：“啊啊不是。”<br>客服：“请问您是在购买什么商品呢？”用户：“呃，龙虾尾。”<br>客服：“那您可以告诉我是看得见摸得着的实物商品，还是虚拟物品呢？”用户：“啊，看得到没得见呢。”<br>客服：“最近兼职刷单类的诈骗很多，你是不是在做任务赚拥金或充值返利之类的呢？”用户：“不是不是。”<br>客服：“您在昨天或更久以前和他有过交易吗？”用户：“啊，没有，我们都是在微信交易，这是在支付宝交易。”，信息过量……<br>客服：“请问您交易的平台名称是什么呢？”用户：“没有平台。”<br>客服：“对方是否提供了商品的交易平台链接呢?”用户：“没有。”…… |
| 期望输出 | 分析 | 在这段对话中，有几个线索表明用户在欺骗。1. 用户回答“啊啊不是”和“不是不是”时，回答的语气模糊，没有直接否认购买商品或做任务赚取佣金的行为。这种回答模糊的方式可能是用户在隐瞒真相。2. 用户回答购买商品是“龙虾尾”，但在被问及商品是实物还是虚拟物品时，用户回答“啊，看得到没得见呢”。这种回答模糊的方式可能是用户在试图混淆客服人员对商品性质的判断。3. 当客服人员问及是否有交易平台链接时，用户回答“没有平台”和“没有”。这种回答可能是用户在试图掩饰交易平台的存在，或者是在隐瞒使用非正规平台进行交易的事实。…… |
| | 标签 | 1 |
| 实际输出 | | 在这段对话中，有几个线索表明用户在欺骗。首先，用户在回答客服关于购买东西的问题时，回答了“啊啊不是”，这种模糊的回答可能是在隐瞒真相。接着，当客服问及用户购买的商品时，用户回答了“龙虾尾”，但这个回答并没有得到后续问题的回应，这可能意味着用户在编造一个虚假的购买故事。此外，用户在被问及商品是否可视可摸时回答了“啊，看得到没得见呢”，这种模棱两可的回答可能是为了掩盖虚拟物品的真实性质。最后，用户回答交易方式为“没有平台”和“没有平台链接”，这种回答可能是在试图掩盖使用非正规平台进行交易的事实。综上所述，以上几个线索表明用户在对话中存在欺骗行为。<br>因此，答案为：是。 |

Figure 5: **FinLIE**测试集中分类正确的样例分析。

| | | |
|---|---|---|
| 输入 | 指令 | 欺骗（英语：Deception）是一种误导他人、隐瞒真相，或者推广假的想法、概念或信念的一种行为或陈述。你的任务是分析客服人员与用户之间的对话，以在对话中搜寻表明用户在欺骗的线索，最终确定用户是否在欺骗。 |
| | 对话 | 客服：“您是如何找到市场调研这份兼职的呢？”用户：“哦一般就是去一些求职的APP然后Marketin去找相关的一些比如呃呃Marketing啊这种。”，说话不连贯<br>客服：“您觉得市场调研的优点有哪些？”用户：“呃优点我觉得呃。呃通过市市场调研确实是1个非常重要的1个点，然后通通过市场调研，能够了你才能你才能够了解现在市场它到底发展的1个概况是怎样的？它发展的1个情况是怎样的，然后呃你才能去。嗯，然后你才能就这样。”，说话不连贯<br>客服：“您觉得是在市场调研中最重要的技能是什么？”用户：“沟通能力吧”<br>客服：“你觉得这份兼职对你就未来的职业规划有什么帮助吗？”用户：“可以提高我的沟通能力，因为因为你是去做市场调查，你不能。非常直接的就是更多的是去市场去做调查，所以自己的这样1个变通和沟通能力，我觉得是非常重要的。”…… |
| 期望输出 | 分析 | 在这段对话中，有几个线索表明用户在欺骗。1. 用户回答“哦一般就是去一些求职的APP然后Marketin去找相关的一些比如呃呃Marketing啊这种”时，回答的不连贯，对于一个简单问题反复强调Marketing。这种回答方式可能是用户在给自己不知道具体途径而找的借口。2. 用户回答市场调研的优点时，很明显说话断断续续，表明用户可能不知道具体的优点，而是随便编造的。…… |
| | 标签 | 1 |
| 实际输出 | | 在这段对话中，我没有找到明显的线索表明用户在欺骗。用户的回答似乎是合理的，没有出现明显的矛盾或不一致之处。用户对市场调研的优点、技能要求以及兼职对未来职业规划的帮助等问题都给出了合理的回答。<br>因此，答案为：否。 |

Figure 6: **IDLIE**测试集中LD-LLM分类错误的样例分析。

图6则展示了LD-LLM分类错误，但是人类分类正确的样例。从图中可以看出LD-LLM对于说话不连贯这一谎言指征不敏感，说明如果上下文没有明显的语义不一致线索，LD-LLM仅利用谎言指征并不能很好地分类。但是人类能可以仅凭谎言指征就能识别谎言，说明在处理谎言识别这类复杂任务，模型更加需要结合谎言指征以及语义不一致线索，而人类可以仅利用其中一种就可以很好的识别谎言，更加表明人类在谎言识别任务上具有不可替代的优势。

## 5 相关工作

### 5.1 谎言识别

在早期的谎言检测阶段，主要依赖于心理学方法(Vrij et al., 2019; Soldner et al., 2019)，这些方法通过分析口头和非口头线索(Granhag et al., 2015)，包括生理功能(Fienberg et al., 2003)等来揭示谎言迹象。Toma(Toma and Hancock, 2010)使用LIWC(Newman et al., 2003; Pennebaker et al., 2015)来识别欺诈性的线上聊天文件；Ott等人(Ott et al., 2011)采用了新的特征来识别具有欺骗性的观点垃圾信息。接着，(Petrov et al., 2006; Soldner et al., 2019)利用句法分析获得的特征来帮助鉴谎。与我们工作较为接近是(Hazra and Majumder, 2024)，其使用大语言模型从语言中发现说谎的线索（如过度自信，模棱两可等），再基于这些线索来判断是否说谎。

### 5.2 事实核查

事实核查分为三个关键步骤(Guo et al., 2022)：主张检测、证据检索和主张验证。主张检测是最初的阶段，旨在辨别需要验证的主张。证据检索需要从各种来源搜集支持或反驳的证据。主张验证采用二元分类，即将主张划分为真实或虚假(Potthast et al., 2017)。事实核查有多种应用，如检测虚假新闻(Chen et al., 2022; Li et al., 2021; Thorne et al., 2018; Wang, 2017)。

### 5.3 微调大模型在特定领域的应用

近年来，大语言模型展现出了惊人的文本理解和生成能力，并在多个自然语言处理任务中取得了优异的表现。为了提高大语言模型对不同领域的适配性，研究者将大语言模型在特定领域的数据上进行微调。这使我们能够将大语言模型应用到广泛的任务场景中，包括信息抽取（InstructUIE (Wang et al., 2023)），写作助手（Writing-Alpaca((Zhang et al., 2023))），医疗咨询（ChatDoctor(Li et al., 2023)），数学算术（Goat(Liu and Low, 2023)），代码生成（WizardCoder(Luo et al., 2023)）等。

## 6 结论

本文专注于研究在真实场景中用户与客服之间对话的谎言识别任务。为了进行更可靠的检测，提出了谎言指征和语义不一致线索的概念作为辅助信息。文章引入了一个谎言识别框架来训练LD-LLM，该框架利用谎言指征并生成不一致线索来确定用户是否在说谎。基于金融风控和身份识别场景中的两个谎言识别数据集**FinLIE**和**IDLIE**，文章构建了一个指令微调数据集，并使用指令微调数据集对LD-LLM进行微调。实验结果表明，谎言识别框架可以在真实场景的自动化谎言识别中达到最优的性能。

## 7 致谢

该论文得到了国家自然科学基金委(No.62176058)和科技部重点研发计划（2023YFF1204800）的项目经费支持。复旦大学CFFF平台为该项目提供了算力支持。

## 参考文献

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Jiangjie Chen, Qiaoben Bao, Changzhi Sun, Xinbo Zhang, Jiaze Chen, Hao Zhou, Yanghua Xiao, and Lei Li. 2022. Loren: Logic-regularized reasoning for interpretable fact verification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10482–10491.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Nadia K Conroy, Victoria L Rubin, and Yimin Chen. 2015. Automatic deception detection: Methods for finding fake news. *Proceedings of the association for information science and technology*, 52(1):1–4.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Stephen E Fienberg, JJ Blascovich, John T Cacioppo, RJ Davidson, P Ekman, DL Faigman, PL Grambsch, PB Imrey, EB Keeler, KB Laskey, et al. 2003. *The polygraph and lie detection*. New York: National Academy Press.

Tommaso Fornaciari and Massimo Poesio. 2013. Automatic deception detection in italian court cases. *Artificial intelligence and law*, 21:303–340.

Pär Anders Granhag, Aldert Vrij, and Bruno Verschuere. 2015. *Detecting deception: Current challenges and cognitive approaches*. John Wiley & Sons.


(c) 2024 中国中文信息学会计算语言学专业委员会

David Gros, Yu Li, and Zhou Yu. 2021a. The R-U-a-robot dataset: Helping avoid chatbot deception by detecting user questions about human or non-human identity. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6999–7013, Online, August. Association for Computational Linguistics.

David Gros, Yu Li, and Zhou Yu. 2021b. The rua-robot dataset: Helping avoid chatbot deception by detecting user questions about human or non-human identity. *arXiv preprint arXiv:2106.02692*.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Sanchaita Hazra and Bodhisattwa Prasad Majumder. 2024. To tell the truth: Language of deception and language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. Association for Computational Linguistics.

Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. 2021. Vln bert: A recurrent vision-and-language bert for navigation. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 1643–1653.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Jinyu Li et al. 2022. Recent advances in end-to-end automatic speech recognition. *APSIPA Transactions on Signal and Information Processing*, 11(1).

Jiawen Li, Shiwen Ni, and Hung-Yu Kao. 2021. Meet the truth: Leverage objective facts and subjective views for interpretable rumor detection. *arXiv preprint arXiv:2107.10747*.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Tiedong Liu and Bryan Kian Hsiang Low. 2023. Goat: Fine-tuned llama outperforms gpt-4 on arithmetic tasks.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Ilya Loshchilov and Frank Hutter. 2018. Fixing weight decay regularization in adam.

Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. 2023. Wizardcoder: Empowering code large language models with evol-instruct.

Matthew L Newman, James W Pennebaker, Diane S Berry, and Jane M Richards. 2003. Lying words: Predicting deception from linguistic styles. *Personality and social psychology bulletin*, 29(5):665–675.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. *arXiv preprint arXiv:1107.4557*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of liwc2015. Technical report.

Denis Peskov and Benny Cheng. 2020. It takes two to lie: One to lie, and one to listen. In *Proceedings of ACL*.

Denis Peskov, Benny Cheng, Ahmed Elgohary, Joe Barrow, Cristian Danescu-Niculescu-Mizil, and Jordan Boyd-Graber. 2020. It takes two to lie: One to lie, and one to listen. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3811–3854, Online, July. Association for Computational Linguistics.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 433–440.


(c) 2024 中国中文信息学会计算语言学专业委员会

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2017. A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638*.

Felix Soldner, Verónica Pérez-Rosas, and Rada Mihalcea. 2019. Box of lies: Multimodal deception detection in dialogues. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1768–1777.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.

Catalina L Toma and Jeffrey T Hancock. 2010. Reading between the lines: linguistic cues to deception in online dating profiles. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 5–8.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Cristian Vaccari and Andrew Chadwick. 2020. Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media+ Society*, 6(1):2056305120903408.

Aldert Vrij, Maria Hartwig, and Pär Anders Granhag. 2019. Reading lies: Nonverbal communication and deception. *Annual review of psychology*, 70:295–317.

Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, Jihua Kang, Jingsheng Yang, Siyuan Li, and Chunsai Du. 2023. Instructuie: Multi-task instruction tuning for unified information extraction.

William Yang Wang. 2017. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Yue Zhang, Leyang Cui, Deng Cai, Xinting Huang, Tao Fang, and Wei Bi. 2023. Multi-task instruction tuning of llama for specific scenarios: A preliminary study on writing assistance.

羊芙葳. 2010. 谎言的识别研究. Ph.D. thesis, 华中科技大学.

CCL 2024

# 基于动态聚类与标签空间映射的上下文学习模板构建方法

**张琦　金醒男　裴誉　杜永萍**
北京工业大学，北京
{ZQi74091,peiyu}@emails.bjut.edu.cn, jinxingnan@outlook.com, ypdu@bjut.edu.cn

## 摘要

面向大语言模型提供自然语言指令，可生成预期输出，体现了其上下文学习能力。上下文学习的性能与上下文模板质量密切相关，现有的工作通常使用单一的选择算法进行模板构建，无法充分激发上下文学习能力。本文提出基于动态聚类与标签空间映射的上下文学习模板构建方法，动态选择相关示例，进一步提出聚类筛选方法，实现不同语义簇中示例多样化的选择。设计基于损失函数的排序选择方法，评估模板学习正确标签空间映射分布的能力，排序形成最终模板。在自然语言推理等任务中的实验结果表明，本文提出的方法使两个不同的大语言模型准确率最高分别提升3.2%和8.9%。

**关键词：** 大语言模型；上下文学习；模板构建；动态聚类

# In-Context Learning Demonstration Construction Method based on Dynamic Clustering and Label Space Mapping

**Qi Zhang　Xingnan Jin　Yu Pei　Yongping Du**
Beijing University of Technology，Beijing, China
{ZQi74091,peiyu}@emails.bjut.edu.cn, jinxingnan@outlook.com, ypdu@bjut.edu.cn

## Abstract

Providing natural language instructions to large language models to generate the expected output reflects the In-Context Learning capability. The performance of In-Context Learning is closely related to the quality of demonstration. Existing methods typically use a single selection algorithm for demonstration construction, which fail to fully exploit the In-Context Learning capability. In this paper, we propose an In-Context Learning demonstration construction method based on dynamic clustering and label space mapping, which dynamically selects relevant samples. Then we propose a clustering filtering method to achieve diversified selection of samples within different semantic clusters. Furthermore, we design a ranking selection approach based on loss functions to assess the demonstration's ability of learning correct mapping distributions in label spaces, and form the final demonstration by sorting samples. Experimental results of different tasks such as natural language inference show that our method improves the accuracy of two large language models by up to 3.2% and 8.9%, respectively.

**Keywords:** Large Language Model , In-Context Learning , Demonstration Construction , Dynamic Clustering

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版
**基金项目**：国家重点研发计划(2023YFB3308004），国家自热科学基金（92267107）

## 1 引言

大语言模型（Large Language Model, LLM）具有庞大的参数规模，在多种任务中展现出强大性能。当参数规模达到一定程度时，LLM表现出了预训练语言模型中没有表现出来的能力，其中，上下文学习（In-Context Learning, ICL）是具有代表性的能力之一(Zhao et al., 2023)。上下文学习是自然语言处理中一种新的范式，LLM只需根据由几个示例组成的上下文模板进行预测，而无需进行任何参数的更新，具有良好的可解释性和可行性(Brown et al., 2020)，因此通过研究上下文学习来提升LLM的能力已经成为新的热点。

现有对上下文学习的研究工作，专注于使用不同的评价指标作为度量，从训练集中选择若干示例，或通过不同的检索方法在大量数据中搜索示例，来构建上下文模板(Dong et al., 2022)。然而，不同测试样例与模板的相关程度不同，使用同一种度量方式构建模板不一定适用于所有样例。此外，上下文模板不仅对示例的要求较高，同时也受示例之间顺序关系的影响，如果不同时考虑这些因素，则会导致上下文学习难以达到理想效果。本文提出的方法相比于已有工作KATE(Liu et al., 2022)，能够根据不同的输入自动调整示例选择，有利于选取优质示例，同时比Levy et al.(2023)的工作使用更少的示例达到理想效果。此外，本文工作能够充分利用训练集中的标签空间映射信息，以此作为标准来判断不同排序方式蕴含正确信息分布的能力，不需要高质量的检索集，相比于Wu et al.(2023)的工作具有更好的泛化性。

本文工作主要贡献如下：

（1）本文提出了动态选择方法，通过对传统Top-K算法进行改进，动态选取与不同测试样例相关的示例，有效剔除无关示例，增加模板与问题的相关性；

（2）本文提出了聚类筛选的方法，采用聚类算法对动态选择后的示例进行划分，并在每个划分后的类中选择示例组成模板，在保证上下文模板示例相关性的同时，增加示例的多样性；

（3）本文针对筛选出的上下文示例，采用交叉熵作为损失函数，选择包含正确标签空间映射的示例排序方式进行排序，提升模板对测试样例标签的预测准确率。

## 2 相关工作

### 2.1 大语言模型与上下文学习

大型语言模型通常包含数千亿参数，是基于Transformer的语言模型，如GPT-3(Brown et al., 2020)、PaLM(Chowdhery et al., 2023)、LLaMA(Touvron et al., 2023)。由于模型规模庞大，计算资源有限，Hoffmann et al.(2022)使用缩放技术进行更加高效的计算资源分配；Rasley et al.(2020)提出了Deepspeed框架，支持分布式训练算法来学习LLM的网络参数。除此之外，提示学习与语言模型的结合，也能够实现在较低资源条件下完成自然语言处理任务。言佳润、鲜于波(2023)提出了基于微调与提示学习的大模型算法，在小样本或零样本数据集中，将GPT与提示学习相结合，较好地完成辩论关系识别任务。穆建媛et al.(2023)基于提示学习，进行中文短文本分类，使模型在少样本情况下取得了性能提升。

以上方法需对模型进行微调，当模型规模变大时，仍需要较多计算资源。上下文学习是LLM出现的一种新兴能力，它只需向大语言模型提供少量示例，即可让模型从中学习到上下文相关知识，并在不更新参数的情况下进行正确推理，目前，上下文学习已被广泛应用于传统自然语言处理任务(Kim et al., 2022)，如机器翻译(Zhu et al., 2023)、信息抽取(Wan et al., 2023)，以及数学推理问题(Wei et al., 2022)。Sun et al.(2022)表明，ICL凭借无需参数更新的学习框架大大降低了对计算资源的要求。此外，已有研究工作表明(Liu et al., 2022)，由于ICL的上下文模板采用自然语言形式，具有较强的可解释性，并且仅调整上下文模板即可将上下文知识与LLM融合。

为了进一步挖掘LLM的上下文学习能力，研究人员进行了大量的工作。Min et al.(2022)提出通过MetaICL来减小预训练和下游任务之间的差距，提高LLM的少样本学习能力。Wei et al.(2023)提出了符号微调技术，利用“输入-标签对”微调LLM，以此促使模型充分挖掘并利用上下文信息中蕴含的语义信息。然而，ICL的性能过度依赖上下文模板的质量。Zhao et al.(2021)研究表明，上下文示例的选择、排列顺序等因素在很大程度上影响着ICL的表现，这激励着研究者进一步探索如何设计一组高质量的上下文学习模板。此外，由于LLM本质上是基于大规模文本语料库训练的文本生成器，因此在分类任务上表现不佳。本文面向NLP领域的七个不同的分类任务数据集，提出基于动态聚类与标签空间映射的上下文学习模板构建方法，增强了LLM的上下文学习能力，提升LLM在分类任务中的表现。


(c) 2024 中国中文信息学会计算语言学专业委员会

### 2.2 上下文学习模板设计

上下文学习的示例模板为LLM提供丰富的语义信息，对上下文学习起着至关重要的作用。通常ICL即是从训练集中选取少量示例，将其改写为自然语言形式，形成上下文示例。随后，将待预测的测试样例与上下文示例连接，形成上下文模板，并将整个模板作为LLM的输入(Dong et al., 2022)。LLM从模板中学习数据与标签之间的映射关系，以及其中蕴含的语义信息，对测试数据进行正确预测。

在给定训练集的情况下，模板设计的重点包括示例子集的选择，以及如何对它们进行合理排序。互信息被用作选择度量(Sorensen et al., 2022), 其优点是不需要标记的样本和特定的最小二乘模型。受到使用检索模块增强神经网络的启发，KATE策略(Liu et al., 2022)被提出，选择与测试样例的语义信息相似的示例构建上下文学习模板，取得了较好的性能。此外，Gonen et al.(2023)发现，提示模板的质量与模型对其所涵盖语言的熟悉程度相关联，并发现模型对模板的困惑度越低，执行任务的能力就越强。Levy et al.(2023)考虑到示例的多样性，提出了一种选择不同上下文示例的方法，旨在覆盖输出内容中的所有结构，鼓励模型从这些结构中学习。在基于检索的方法中，Rubin et al.(2022)提出了一个二阶段的检索方法来选择示例，使用标记数据训练一个检索器，用于检索示例并构建模板。为了解决先前工作难以在各种不同任务之间迁移的问题，Li et al.(2023)提出多任务统一检索器，充分融合各类任务相关知识，使模型取得了优秀的性能表现。本文工作优化传统Top-k算法，根据不同输入样例动态选择上下文示例，增加示例与测试样例之间的相关性，并使用聚类策略提升示例的多样性。

除示例的选择外，对选定的示例进行合理排序也是模板设计的重点任务之一。研究表明，不同的语言模型均存在对示例排序敏感这一问题(Lu et al., 2022)。现有工作通常通过计算示例与测试样例之间的负欧几里得距离或余弦相似度，并以此进行排序(Liu et al., 2022)。熵值同样被用作度量指标，可分为全局熵和局部熵度量，并证实熵度量和ICL性能之间存在正相关关系(Lu et al., 2022)。Wu et al.(2023)受Solomonoff(1964)的一般推理理论和信息论的最小描述长度原则的启发，设计了一种基于最小描述长度的排序算法。本文基于输入数据与标签之间的映射关系，设计损失函数进行度量，为LLM提供充分的标签映射信息，提升预测准确率。

## 3 方法

上下文学习模板设计旨在通过对训练集中标注数据进行筛选、排序等操作，将其与测试样例相连接构建示例模板。具体可描述为：假设有上下文标注示例$\mathcal{D}$,$\mathcal{D} = [D_1, D_2, \ldots\ldots, D_i, \ldots\ldots D_j]$，其中$D_i = T(x_i, y_i)$，T代表将数据转换为自然语言的操作，$x_i$为第i个样本的输入，$y_i$是$x_i$的标签。对于给定测试样本$x'$，使用LLM预测对应标签$y'$可表示为：

$$y' = argmax_{y \in Y} \frac{P(y|\mathcal{D} \oplus x')}{\sum_{y \in Y} P(y|\mathcal{D} \oplus x')} \tag{1}$$

基于动态聚类与标签空间映射的上下文学习模板构建方法整体结构如图1所示。首先采用动态搜索方法，对测试样例进行实例级的上下文示例的选择，进一步设计聚类算法，采用两种不同的样本抽取方式择优筛选，确定最终的上下文示例集，最后，根据上下文模板中蕴含的标签空间映射关系分布进行示例排序，增强模板的有效性。

### 3.1 动态选择

传统Top-K算法可能会导致选取的示例与测试样例的语义关系不匹配，或遗漏掉语义关系相近的优质示例。有研究表明，可通过设置参数对距离进行筛选，减少预测误差(张少辉et al., 2022)。基于该发现，本方法改进传统Top-K算法，根据测试样例与其他示例之间的欧式距离来动态选取候选示例。欧式距离的计算如公式2所示：

$$L_2(x_i, x_j) = \sqrt{\left(\sum_{j=1}^{n} |x_i^{(1)} - x_j^{(2)}|^2\right)} \tag{2}$$

本方法首先使用all-mpnet-base-v2模型中的编码器，将数据中的句子和段落映射到768维的密集向量空间，形成大小为(64,768)的多维矩阵，并创建Faiss索引，使用内积计算，得到同样


(c) 2024 中国中文信息学会计算语言学专业委员会

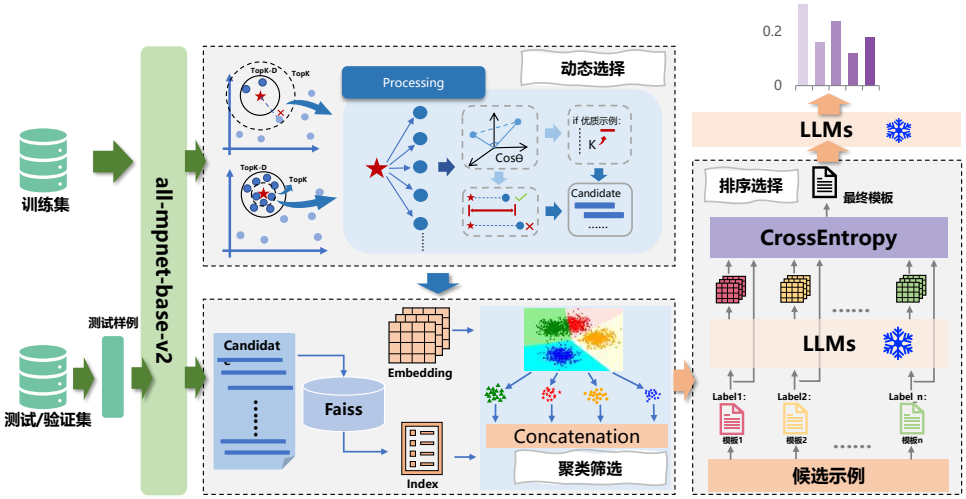



Figure 1: 基于动态聚类与标签空间映射的上下文模板构建方法

为768维的索引，将嵌入向量添加到Faiss索引中，完成样本的编码操作。随后获取测试样例$x_i$和其他样本$x_j$之间的欧氏距离，并由小到大进行排序。再根据K值筛选上下文示例，根据示例与测试样例之间的距离设置距离阈值，作为判断示例质量的依据。为了提升算法的准确度，将阈值分为优质阈值和无关阈值。当该组示例中有超过半数的数据与测试样例之间的距离小于无关阈值，则认为测试数据周围的相似样本稀疏，而当示例中距离最远的数据仍然大于优质阈值，则认为测试数据周围的相似样本稠密。

当测试样本周围的相似示例稠密，则动态扩大K值的范围；当测试样本周围的相似示例稀疏，则剔除与测试样本之间距离大于无关阈值的样本。为了防止剔除无关样本后候选样本数量过少，本方法随时动态监测上下文示例的数量，并设置上下文模板示例数量的上限和下限参数。根据欧氏距离选取数量等于上限参数的相关样本，从而保证候选示例与测试样本的相关性。

### 3.2 聚类筛选

动态选择模块对上下文示例进行了初步筛选，保证了模板中示例之间的相关性。然而，若直接对动态选择模块输出的候选示例进行抽取并排列，搜索空间为指数级，这将消耗大量的时间和计算资源。此外，初步筛选后的示例集中蕴含丰富多样的上下文语义信息，若从中随机抽取少量示例作为模板，存在知识信息冗余、提示角度片面等问题。例如：在Trec主题分类数据集中，当输入数据为“What class of animals makes up more than two-thirds of known species?”时，一个良好的上下文模板应该从“animals”、“species”、“ratio”等角度考虑问题。为了最大限度地挖掘初筛后示例集的信息，该模块使用KMeans聚类算法，计算动态选择后的候选示例在映射空间中的距离关系并进行分簇。经过聚类筛选后的上下文示例被分为若干簇，每个簇内部的示例具有相似的语义信息，而不同簇中示例的语义信息差别较大。从每个簇中进行示例抽取构建模板，充分地利用上下文示例中不同簇的语义信息，促使大语言模型从多角度进行综合决策。在对每个簇进行示例抽样时，为了同时考虑上下文示例的相关性和多样性，采用随机抽取和顺序抽取两种不同方法，如图2所示。

**随机抽取：**从每个上下文示例簇中随机抽取少量示例组成上下文模板的候选示例集，该方法保证了示例多样性。

**顺序抽取：**基于每个簇中示例与测试样本之间在嵌入空间的距离，依次从每个簇中顺序抽取相关度高的示例组成候选示例集，保证示例的高相关性。


(c) 2024 中国中文信息学会计算语言学专业委员会

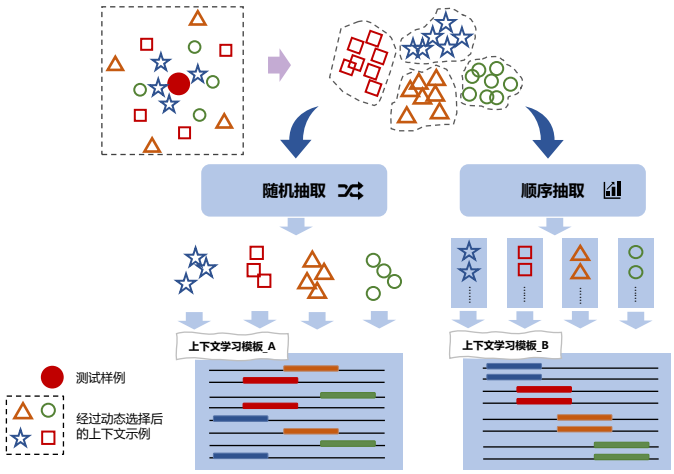



Figure 2: 聚类筛选过程

### 3.3 排序选择

动态选择和聚类筛选过程从上下文示例的相关性与多样性的角度筛选候选示例，而没有考虑示例的排列顺序对模板质量的影响。然而，由于LLM对于上下文示例的顺序较为敏感，所以对上下文示例进行合理排序是至关重要的。受Min et al.(2022)和高怡et al.(2023)的启发，ICL的性能收益主要来自于独立规范的输入空间和标签空间，以及正确一致的"输入-标签对"的格式，且模型可以通过模板来引入标签信息，因此我们考虑输入空间与标签空间之间的映射关系，为上下文示例选择合理的排序。

首先需要确定待筛选的不同排序方式。对于动态选择和聚类筛选操作后的N个候选示例，具有N!种不同的排列方式。为了节省计算资源，本方法进一步减少示例数目，从N个候选示例中选择n个示例，并进行排列。考虑到时间和计算资源的复杂度，本方法从所有可能的排列空间中随机选取了m种排列进行比较，如公式3所示：

$$R = Random_m\left(\frac{N!}{(N-n)!}\right), n \leq N \tag{3}$$

其中R代表包括m种不同排序的集合，每种排序都包含n个示例，$Random_m$的含义是从n!种排列方式中选取m个。

由于在预测前无法获取测试样例的真实标签，本方法获取其所有可能的标签空间信息，将输入数据的每一个可能的标签记为c。每一种可能的标签c对应多种不同的排序方式r，对于使用排序方式r进行排序的上下文示例文本序列$T_{c,r}$，使用分词器中的编码器对模板进行编码，编码后获得包含正确映射关系的信息分布，记作$emb_{T_{c,r}}$，将$T_{c,r}$与输入样例$x_t$拼接后，作为模板提示LLM输出预测结果。本文获取预测结果的嵌入空间信息分布，记作$emb_{T_{c,r},x_t}$。由于$T_{c,r}$由带标签的训练数据组成，其编码后具有正确的输入空间和标签空间之间的映射信息分布，因此可以将该映射信息分布作为标准，来判断该排序方式能否为LLM提供正确信息。本方法使用交叉熵损失函数为每种排序方式进行打分，如公式4所示。其中，$\mathcal{C}$代表分类类别标签的数目。将得到的损失值进行归一化后，选取损失最小的排序序列作为最终模板。

$$L(emb_{T_{c,r}}, emb_{T_{c,r},x_t}) = -\sum_{c=1}^{\mathcal{C}} emb_{T_{c,r}} log(emb_{T_{c,r},x_t}) \tag{4}$$

## 4 实验

### 4.1 数据集与实验设置

为了验证基于动态聚类与标签空间映射的上下文学习模板构建方法的有效性，本文选取七项自然语言处理领域的分类任务数据集进行实验，包括情感分析数据集SST-2、SST-5，自然语言推理数据集SNLI、MNLI、QNLI，以及主题分类任务数据集Trec、Ag_News。


(c) 2024 中国中文信息学会计算语言学专业委员会

本实验环境为CPU配置为Intel(R) Core(TM) i9-10980XE CPU@3.00GHz，GPU配置为NVIDIA GeForce RTX 3090的服务器，CUDA版本为12.2，使用Pytorch v2.1.0作为深度学习框架。由于本实验以实例级别进行，不同实例对应的上下文模板长度不同，因此进行推理时将batch size设为1。在实验过程中，将随机数种子分别设为1,2,3，选取最优结果。

### 4.2 对比实验

本文基于大语言模型GPT2-XL和Bloom，将基于动态聚类与标签空间映射的上下文学习模板构建方法与其他性能优秀的方法进行对比，在不同数据集上进行实验，实验结果如表1所示。

- Random：该方法采用随机选取的策略，为每一个数据集构建一个上下文模板。
- Prompt：该方法属于上下文学习的特殊情况，即不使用任何示例，直接进行预测。
- MI：(Sorensen et al., 2022)该方法选择最大化输入和输出之间的互信息进行模板构建，无需标记示例，也不需要使用模型进行大规模推理预测。
- GlobalE：(Lu et al., 2022)为了避免类别不平衡的预测问题，该方法使用局部熵值作为评价指标，为上下文模板选择合适的示例排列顺序。
- Top-K+LocalE：(Lu et al., 2022)该方法根据模型对输入的预测结果的偏移程度，判断模型的质量以及模型在不同类别之间进行分类的能力。使用Top-k方法进行示例选择后，进一步使用LocalE方法进行示例间排序。
- KATE：(Liu et al., 2022)该方法使用KNN算法辅助示例选择，并使用编码器将样例映射到语义空间中，检索语义上与测试样例相似的示例，构建与其相应的模板。

| Model | Method | Dataset | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | SST2 | SST5 | SNLI | MNLI | QNLI | Trec | Ag_News |
| GPT2-XL(1.5B) | Random | 31.47 | 26.24 | 44.75 | 40.98 | 52.35 | 19.40 | 33.07 |
| | Prompt | 47.12 | 29.46 | 41.98 | 39.06 | 50.45 | 13.80 | 29.76 |
| | MI | 52.86 | 35.35 | 46.02 | 41.32 | 50.62 | 16.00 | 47.29 |
| | GlobalE | 67.27 | 33.21 | 46.99 | 40.46 | 57.27 | 28.53 | 52.01 |
| | Top-K+LocalE | 63.21 | 34.03 | 51.71 | 42.46 | 53.45 | 35.80 | 83.42 |
| | KATE | 68.75 | 36.97 | 58.34 | 46.33 | 59.77 | 40.80 | 88.87 |
| | **Ours** | **71.94** | **39.28** | **59.57** | **47.15** | **61.16** | **43.00** | **89.21** |
| Bloom(1.1B) | Random | 34.87 | 22.31 | 43.39 | 41.87 | 49.65 | 18.00 | 32.95 |
| | Prompt | 25.04 | 26.02 | 42.26 | 39.85 | 49.61 | 5.80 | 25.13 |
| | Top-K+LocalE | 54.31 | 29.70 | 51.52 | 37.95 | 48.95 | 28.20 | 75.41 |
| | KATE | 62.49 | 35.84 | **58.74** | 43.37 | 54.99 | 39.40 | 85.50 |
| | **Ours** | **71.39** | **36.97** | 57.85 | **44.12** | **55.41** | **39.60** | **86.82** |

Table 1: 不同上下文模板组织方法在不同任务数据集上的性能对比

从表1的数据中可以看出，本文所提出的基于动态聚类与标签空间映射的上下文学习模板构建方法，相比于其他基于熵值、互信息等方法，在不同数据集、不同LLM上普遍具有优秀性能表现。其中，在SST-2数据集上，GPT2-XL模型准确率提升了3.19%，Bloom模型准确率提升了8.9%，验证了本文提出方法的有效性。本方法选取了比传统Top-K方法更多的相关样本，并过滤掉了无关样本，减少了干扰信息；利用聚类筛选增加了上下文模板中示例的多样性，使模型能够从多角度进行综合预测；合理选择示例排序方式来构建最终的上下文模板，为模型预测提供了正确的输入与标签之间的映射关系，最终提升了模型预测的准确率。

### 4.3 消融实验

我们使用GPT2-XL模型，分别验证本文提出的动态选择、聚类筛选和排序选择的有效性。


(c) 2024 中国中文信息学会计算语言学专业委员会

### 4.3.1 动态选择方法消融实验

图3展示了在七个数据集上针对动态选择方法的消融实验结果，Top-k代表根据欧式距离进行示例选择的传统方法。Topk-D使用本文提出的动态选择方法改进模板。结果显示，在不同数据集上，动态选择方法能够普遍提升模型预测准确率，说明其可以有效评估上下文模板示例与测试样本之间的相关度。而对于Ag_News数据集，由于其包含四种不同标签，映射关系与信息分布较为复杂。此外，不同于其他输入长度较短的数据集，Ag_News数据集的输入长度集中在200-400之间，对于较长文本中的语义信息，以及句子与标签之间的映射关系，相对更难获取，因此还需对示例进行进一步地筛选和排序。

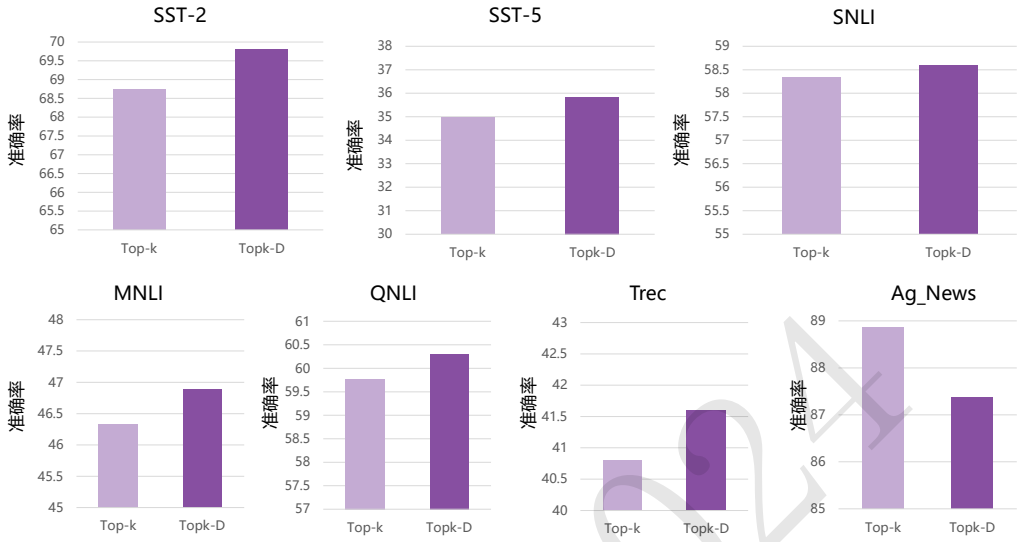



Figure 3: 动态选择方法的对比实验结果

### 4.3.2 聚类筛选方法消融实验

图4展示了针对聚类筛选方法的消融实验结果。其中“TopK-D+顺序抽取”首先使用TopK-D动态选取示例，随后采用聚类筛选方法中的顺序抽取策略得到示例集；“TopK+随机抽取”指在TopK-D算法动态选取示例后，使用聚类筛选方法中的随机抽取策略得到示例集。




Figure 4: 聚类筛选方法的消融实验结果

第二十三届中国计算语言学大会论文集，第883页-第893页，太原，中国，2024年7月25日至28日。
卷1：主会论文
(c) 2024 中国中文信息学会计算语言学专业委员会

### 4.3.3 排序选择方法消融实验

图5是针对排序选择模块进行的消融实验结果。其中“TopK-D+聚类筛选+排序选择”代表先使用TopK-D和聚类筛选算法进行示例选择，随后对示例进行排序选择。从图中的实验结果可以看出，本文提出的排序选择方法能够为LLM提供正确且充分的标签映射关系，在不同数据集上的准确率均得到有效提升，对于标签空间中标签较多、较复杂的数据集，如SST-5、Trec，本文的排序选择方法获得了更为显著的提升效果。

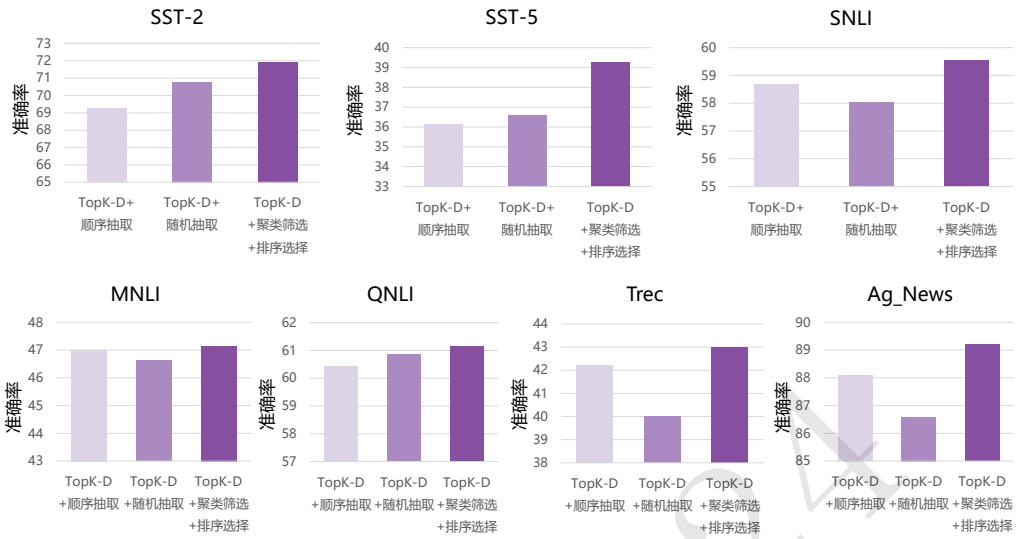



Figure 5: 排序选择的消融实验结果

### 4.4 超参数实验

为了探索不同超参数的设置对模型性能的影响，我们在Trec数据集上，使用GPT2-XL模型对动态选择模块进行了一系列超参数实验，如图6所示。

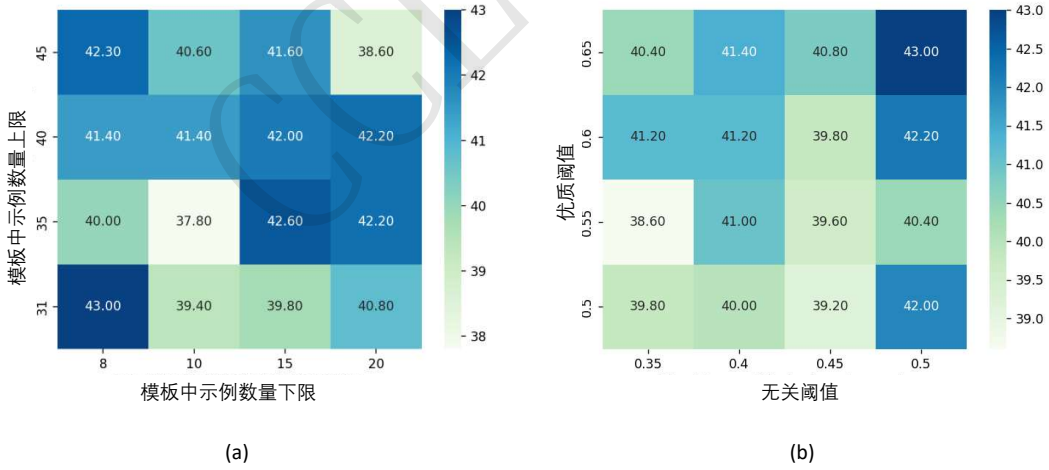



Figure 6: 动态选择模块的超参数实验结果

图6所示的热力图分别展示了动态选择模块中，模板示例数量的上限和下限、优质阈值和无关阈值的取值对性能的影响。图6(a)中结果显示，当模板中示例数量下限为8，上限为31时，LLM达到最优性能43%。当模板示例的数量下限提升时，准确率均有所下降，这说明当嵌入空间的测试样本周围稀疏时，向上下文模板中增加过多示例，存在引入无关示例的风险，会对LLM产生干扰；当模板示例的数量上限提升时，由于此时嵌入空间中的测试样本周


(c) 2024 中国中文信息学会计算语言学专业委员会

围稠密，相关示例较多，因此会提高LLM预测的准确率，但若此时降低示例的下限，提升上限同样容易引入无关示例，会对标签预测造成干扰。因此综合考虑示例数量的上限和下限，本文选择上限为31，下限为8的超参数组合进行动态选择，使用较少的数据量达到较高的性能。从图6(b)中显示的结果可以看出，当无关阈值取0.5，优质阈值取0.65时，动态选择模块能够使LLM达到43%的最优性能。当使用更小的无关阈值或更大的优质阈值时，准确率普遍有所下降，说明其他阈值超参数的组合对示例与测试样例之间距离的判断无法达到理想的示例动态筛选效果。

### 4.5 实例对比

为了分析本文提出方法的有效性，从Trec数据集中选取了主题分类任务的实例数据进行对比，如表2所示。

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 输入<br>分类标签 | What is the full form of .com?<br>ABBR | | | | | | | |
| Top-K<br>预测标签 | ABBR<br>ENTY | ABBR | ENTY | ABBR | ENTY | ENTY | ENTY | ABBR |
| Ours<br>预测标签 | ENTY<br>ABBR | ABBR | ABBR | ABBR | LOC | ENTY | ABBR | ABBR |

Table 2: 本文方法(Ours)与Top-K方法所构建的模板和输出结果的对比，模板均由八个候选示例排列组成。其中ABBR代表与缩写相关的主题，ENTY代表与实体相关的主题，LOC代表与位置相关的主题

从表2的示例数据中可以看出，本文所提出的方法相比于Top-K方法，包含了更多与标签相同的映射分布，保证了预测的准确率，同时也更具有多样性，表明本文方法能够有效提示LLM做出正确预测。

## 5 总结

本文面向大语言模型中上下文学习模板的设计与组织这一任务，提出了基于动态聚类与标签空间映射的上下文学习模板构建方法。采用动态选择策略对Top-K算法进行优化，根据不同的测试样例设置优质阈值和无关阈值，以实例级别进行上下文示例的选择，保证模板的相关性；进一步提出了聚类筛选的方法，利用聚类算法将动态选择后的示例进行再次筛选，在保证示例相关性的前提下，增加上下文学习示例的多样性；提出基于映射损失的排序选择方法，选择包含正确映射关系的示例排序方式构建最终模板。实验在情感分析、自然语言推理和主题分类领域中的七个数据集中进行验证，并针对每个模块进行了充分的消融实验，结果表明本文提出的方法可以为模型提供高质量的上下文学习模板，有效提升LLM在不同任务上的准确率。未来的工作将关注如何将外部知识与上下文模板相融合，并提升LLM的鲁棒性，使其在使用不同上下文模板的条件下达到理想的效果。

## 参考文献

高怡, 纪焘, 吴苑斌, 牟小峰, and 王綻. 2023. 基于标签增强和对比学习的鲁棒小样本事件检测. 中文信息学报, 37(4): 98-108.

穆建媛, 朱毅, 周鑫柯, 李云, 强继朋, and 袁运浩. 2023. 基于提示学习的中文短文本分类方法. 中文信息学报, 37(7): 82–90.

言佳润and 鲜于波. 2023. 面向中文网络对话文本的论辩挖掘——基于微调与提示学习的大模型算法. 中文信息学报, 37(10): 139-148.

张少辉, 亓玉浩, 翟方文, 吕洪波, and 宋亦旭. 2022. 基于冗余距离筛选的UWB定位优化方法. 清华大学学报（自然科学版）, 62(5): 934-942.

Tom B. Brown, Benjamin Mann, Nick Ryder, et al. 2020. Language models are few-shot learners. Advances in neural information processing systems,2020,33:1877-1901.


(c) 2024 中国中文信息学会计算语言学专业委员会

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, et al. 2023. Palm: Scaling language modeling with pathways. Journal of Machine Learning Research, 2023, 24(240): 1-113.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2022. A survey for in-context learning. arXiv preprint arXiv:2301.00234.

Hila Gonen, Srini Iyer, Terra Blevins, Noah Smith, and Luke Zettlemoyer. 2023. Demystifying Prompts in Language Models via Perplexity Estimation. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 10136–10148, Singapore. Association for Computational Linguistics.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models. arXiv preprint arXiv:2203.15556.

Hyuhng Joon Kim, Hyunsoo Cho, Junyeob Kim, Taeuk Kim, Kang Min Yoo, and Sang-goo Lee. 2022. Self-generated in-context learning: Leveraging auto-regressive language models as a demonstration generator. arXiv preprint arXiv:2206.08082.

Itay Levy, Ben Bogin, and Jonathan Berant. 2023. Diverse Demonstrations Improve In-context Compositional Generalization. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1401–1422, Toronto, Canada. Association for Computational Linguistics.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What Makes Good In-Context Examples for GPT-3?. In Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Xiaonan Li, Kai Lv, Hang Yan, Tianyang Lin, Wei Zhu, Yuan Ni, Guotong Xie, Xiaoling Wang, and Xipeng Qiu. 2023. Unified Demonstration Retriever for In-Context Learning. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4644–4668, Toronto, Canada. Association for Computational Linguistics.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022. MetaICL: Learning to Learn In Context. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2791–2809, Seattle, United States. Association for Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020: 3505-3506.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning To Retrieve Prompts for In-Context Learning. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2655–2671, Seattle, United States. Association for Computational Linguistics.

R.J.Solomonoff. 1964. A formal theory of inductive inference. Information and Control, 7(1).

Taylor Sorensen, Joshua Robinson, Christopher Rytting, Alexander Shaw, Kyle Rogers, Alexia Delorey, Mahmoud Khalil, Nancy Fulda, and David Wingate. 2022. An Information-theoretic Approach to Prompt Engineering Without Ground Truth Labels. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 819–862, Dublin, Ireland. Association for Computational Linguistics.


(c) 2024 中国中文信息学会计算语言学专业委员会

Tianxiang Sun, Yunfan Shao, Hong Qian, Xuanjing Huang, and Xipeng Qiu. 2022. Black-box tuning for language-model-as-a-service. International Conference on Machine Learning, PMLR.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying Liu, Haiyue Song, Jiwei Li, and Sadao Kurohashi. 2023. GPT-RE: In-context Learning for Relation Extraction using Large Language Models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 3534–3547, Singapore. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems 35 (2022): 24824-24837.

Jerry Wei, Le Hou, Andrew Lampinen, Xiangning Chen, Da Huang, Yi Tay, Xinyun Chen, Yifeng Lu, Denny Zhou, Tengyu Ma, and Quoc Le. 2023. Symbol tuning improves in-context learning in language models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 968–979, Singapore. Association for Computational Linguistics.

Zhiyong Wu, Yaoxiang Wang, Jiacheng Ye, and Lingpeng Kong. 2023. Self-Adaptive In-Context Learning: An Information Compression Perspective for In-Context Example Selection and Ordering. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1423–1436, Toronto, Canada. Association for Computational Linguistics.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. International Conference on Machine Learning. PMLR.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. arXiv preprint arXiv:2303.18223.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Multilingual machine translation with large language models: Empirical results and analysis. arXiv preprint arXiv:2304.04675.

CCL 2024

# 基于领域信息分解式学习的大语言模型修辞认知增强方法

**王雯[1], 于东[1,*], 刘鹏远[1,2]**
1.北京语言大学信息科学学院，北京，100083
2.国家语言资源监测与研究平面媒体中心，北京，100083
wangwenblcu@gmail.com, yudong@blcu.edu.cn,liupengyuan@blcu.edu.cn

## 摘要

中文修辞手法多样且概念差异性大，大语言模型对部分修辞手法的认知存在缺陷。针对该问题，本文研究如何增强大语言模型的修辞认知能力，并探究其与修辞识别性能之间的关系。为此，本文提出了QAKAG框架，此框架首先引入信息分解式学习思想，通过问答形式检测大语言模型的修辞认知缺陷，然后以四种不同的知识组合方式探究最优信息补充机制，实现了大语言模型修辞认知能力的增强。本文构建了多类别中文修辞句数据集MCRSD和修辞知识库MCRKB，并在ChatGPT4等六个大语言模型上开展实验研究，验证了QAKAG框架对增强大语言模型修辞认知能力的有效性以及其各阶段的必要性。结果表明，在QAKAG框架的增强下，六个大语言模型在多类别修辞识别任务上的性能相较直接回答识别问题的平均F1值提高22.1%，优于Zero-shot-CoT、RAG-BaiKe、Few-Shot5提示策略。

**关键词：** 大语言模型；修辞认知；分解式学习；动态信息补充

# Method for Enhancing Rhetorical Cognition of Large Language Models Based on Decomposed Learning of Field Information

**Wen Wang[1],Dong Yu[1,*], Pengyuan Liu[1,2]**
1.Faculty of Computer Science,Beijing Language and Culture University,Beijing,100083
2.National Language Resources Monitoring and Research Center for Print Media,Beijing,100083
wangwenblcu@gmail.com, yudong@blcu.edu.cn,liupengyuan@blcu.edu.cn

## Abstract

Chinese rhetorical devices are diverse and conceptually different, weakening large language models in the cognition of some rhetorical devices. This paper aims to enhance the large language model's rhetorical cognition ability and explore the relationship between ability and performance. Therefore, the paper proposes the QAKAG framework, which firstly applies the idea of information decomposition learning and detects the rhetorical cognitive deficiencies of the large language model by asking it questions, and then explores the optimal information supplementation mechanism out of four knowledge combination methods, which enhances rhetorical cognitive ability. Based on a multi-category Chinese rhetorical sentence dataset (MCRSD) and a multi-category

*为通讯作者
**基金项目：** 教育部人文社科规划项目（23YJAZH184）；北京语言大学梧桐创新平台(中央高校基本科研业务费)(21PT04)；北京语言大学研究生创新基金（中央高校基本科研业务费专项资金）项目成果（24YCX114）
©2024 中国计算语言学大会根据《Creative Commons Attribution 4.0 International License》许可出版

rhetorical knowledge base (MCRKB), we carried out experimental studies on six large language models, including ChatGPT4, to prove the effect of QAKAG framework enhancing the rhetorical cognitive ability and the indispensability of its two stages. The results show that, in average, the QAKAG framework improves the performance on multi-category rhetorical recognition tasks by 22.1% measured with F1 value, outperforming the Zero-shot-CoT, RAG-BaiKe, and Few-Shot5 prompting strategies.

**Keywords:** Large Language Models , Rhetorical Cognition , Decompositional Learning , Dynamic Information Supplement

## 1 引言

修辞识别是一个复杂且具有挑战性的任务，旨在探究和理解中文文本中的修辞结构和语义信息，从而揭示其中的情感、态度和意图，为语言理解、信息检索和文本生成等任务提供更深层次的支持和指导(刘广辉,2022)。修辞是用于各种文章或应用文在写作时语言表达方法的集合，在文本中呈现出复杂性和多样性，具有形式灵活、不易量化的特点(王荣文(2018)，修辞识别任务的关键在于对不同修辞手法进行深入理解与认知。通过在海量数据上进行大规模预训练，大语言模型（Large Language Model，LLM）具有了强大的知识表征能力和理解推理能力(崔亿萍,2023)。但在特定领域任务中，大语言模型可能缺乏某些对应的专业知识，导致在解决特定领域问题上表现不佳(Zhang et al., 2023)，而且模型的知识幻觉问题也较为突出(Li et al., 2024b)。因此，本文旨在多类别修辞识别任务上探索如何有效增强大语言模型修辞认知能力，并探究大语言模型修辞认知能力和修辞识别能力之间的关系。

修辞认知，指模型对修辞语言概念的理解、分析和运用能力(徐盛桓,2008)。模型的修辞认知水平与修辞识别能力存在密切相关性，如果模型对某种修辞认知水平较高，其修辞识别能力可能也会更强(邱文生,2015)。传统的识别方法不具备修辞认知能力，将修辞识别任务视为语言层面的分类问题，主要依赖表层特征进行分类，采用基于统计的方法识别比喻、排比等高频修辞(朱晓亮et al., 2019; 呼啸et al., 2020)，对修辞知识缺乏深层次的认知。受限于人工标注成本，修辞识别任务的训练数据集往往规模有限且数据分布不均衡，导致模型识别高频修辞的能力高于低频修辞，且由于数据稀缺问题，对通感、顶真等低频修辞缺乏深入的研究。大语言模型也存在类似问题。通过在小规模修辞语料上测试发现，大语言模型对比喻、排比等高频修辞手法的认知能力较强，但对反复、顶真等低频上的认知能力则有所不足，由此推测大语言模型在高低频修辞手法上的理解存在认知级差。图1展示了ChatGPT3.5(Dai et al., 2022)对比喻和顶真两种修辞的识别情况。可以看到，ChatGPT3.5对比喻的识别和理解能力比较出色，但对顶真缺乏基础认知，导致识别效果较差，说明大语言模型的修辞识别能力仍然是基于训练语料统计特征，对低频修辞的认知能力有所欠缺。




图 1: ChatGPT3.5对比喻和顶真修辞的识别理解情况

因此，本文深入探究了大语言模型在不同修辞手法上的认知表现，并在序列求解思想的启发下，提出了一种基于领域信息分解式学习的QAKAG（**Q**uestion **A**nswering with **K**nowledge **A**ugmented **G**eneration）框架来增强大语言模型的修辞认知能力，缩小大语言模型对于不同修辞手法的认知差异。在QAKAG框架中，首先是领域信息认知缺陷检测阶段，该阶段将测试问题分解成系列子问题，以子问题针对性问答的形式探究大语言模型对特定修辞手法


(c) 2024 中国中文信息学会计算语言学专业委员会

的理解情况。然后是领域信息动态补充阶段，该阶段根据前一阶段对子问题的理解情况，从修辞知识库中为大语言模型动态抽取并补充其基本理解和理解欠缺的修辞知识点。该框架结合了分解式学习的思想和知识增强生成模型的方式，可以提高大语言模型的修辞认知能力。为验证QAKAG框架的有效性，本文构建了多类别中文修辞句数据集MCRSD(**M**ulti-category **C**hinese **R**hetorical **S**entence **D**ataset)和修辞知识库MCRKB(**M**ulti-category **C**hinese **R**hetorical **K**nowledge **B**ase)，并设计了四种不同的知识组合方式进行实验探究。结果表明，QAKAG框架可以有效增强大语言模型修辞认知能力，能够为大语言模型解决复杂认知问题提供参考。

本文的贡献主要包括以下三个方面：第一，提出了QAKAG框架，能够有效增强大语言模型的修辞认知能力，提高其在修辞识别任务上的性能。第二，构建了多类别中文修辞句数据集MCRSD和修辞知识库MCRSB，为验证QAKAG框架的有效性提供了数据基础，也丰富了中文修辞识别研究领域的语言资源。第三，探究了多个大语言模型在不同修辞手法上的认知表现，并对各大语言模型对不同修辞的认知情况进行了详细分析，为大语言模型进一步提升修辞识别能力提供了新的思路和方法。

## 2 相关工作

修辞识别作为自然语言处理中的重要任务，其目标是理解和分析文本中所使用的修辞手法，如比喻、比拟、排比等，具有广泛的应用场景。早期研究集中在修辞学的理论层面，Mann and Thompson (1988)提出了著名的修辞结构理论（RST），为文本修辞的分析和生成提供了理论框架。在20世纪以后，修辞学的研究逐渐扩展到了更广泛的领域，研究者们开始尝试使用各种自然语言处理技术来自动化识别和分析修辞。传统的修辞识别研究主要依赖于语言规则和特征提取。例如国外的PEG(Page,1966)、IEA(Landauer et al., 2000)、E-rater(Attali and Burstein,2004)系统利用多元线性回归、统计学等方法进行特征工程并构建分类器对学生作文进行自动打分；Tsvetkov et al. (2014)提出一种基于语义特征的分类器进行隐喻识别研究。国内学者刘明杨(2015)提出利用排比标检验的启发式方法识别排比和比喻；李明峰and 贾修一(2018)则是采用多分类器集成学习策略进行中文反语识别等。然而，由于修辞的复杂多样性，基于规则和特征的方法往往面临泛化能力不足的问题。随着深度学习的兴起，基于神经网络的模型开始主导修辞识别的研究。长短时记忆（LSTM）、卷积神经网络（CNN）和递归神经网络（RNN）等被广泛应用于此任务中(付瑞吉et al., 2018; 石昀东,2019; 赵琳玲,2020)，但测试数据集规模小，难以覆盖语言的各种变体和复杂构造，限制了修辞识别的准确性和可扩展性。可以看出，传统修辞识别研究关注重点仍在于如何有效提取不同修辞手法的特征进行分类，缺少从修辞认知的角度探讨模型是否对修辞手法的深层含义具有理解能力。

具有强大语言分析理解能力的ChatGPT等大语言模型的出现，为修辞识别任务提供了新的研究途径。Ziems et al.(2024)在零样本设置下评估大语言模型，发现模型在比喻语言分类上表现出色。李玲玲and 王晓燕(2023)在对传统机器翻译和ChatGPT翻译分析对比中发现大语言模型在隐喻理解层面具备显著优势。蔡越(2024)认为可以利用大语言模型生成各种修辞手法的文本，帮助写作者理解和掌握修辞手法。Li et al.(2024a)提出的StyleChat框架可以从词汇、句法、修辞手法、修辞目的四个方面增强大语言模型风格化对话生成能力，侧面说明了大语言模型在理解和应用修辞方面存在潜力。但目前，关注如何利用大语言模型处理修辞识别任务的相关研究还比较少，缺乏对大语言修辞认识能力和修辞识别性能之间的探究分析。

如何增强大语言模型解决复杂问题的能力是近年来的研究热点。Zhou et al.(2022)提出了least-to-most prompting以解决复杂推理问题，主要思想是将一个复杂的问题分解为一系列更简单的子问题依次解决，然后用先前解决的子问题的答案来促进当前子问题的解决。Trivedi et al.(2022)在多步骤QA任务上，用CoT指导检索，再反过来用检索的结果来改进CoT。Liu et al.(2024)提出的RA-ISF框架中的问题分解模块则是可将复杂问题拆解为子问题并迭代处理，这种自反馈机制可以改善大语言模型在复杂任务中的表现。因此，本文将在现有主流的六个大语言模型上针对比喻、比拟、通感、排比、对偶、反复、设问、夸张、反问，顶真十种常用修辞开展认知能力测试，并基于序列求解的思想针对其中的不足提出解决方案，旨在有效提高大语言模型的修辞认知水平，进而提升其在修辞识别任务上的性能。


(c) 2024 中国中文信息学会计算语言学专业委员会

## 3 大模型修辞认知能力评测数据集建设

### 3.1 多类别中文修辞句数据集MCRSD构建

修辞数据集是修辞识别任务的基础。目前公开可用的中文修辞数据集有CMC中文比喻语料库(Li et al., 2022a)、C-SEM语义评测基准基础修饰知识检测（SLRFC）子评测项数据集[0]和CCL 2024发布的“中小学作文修辞识别与理解评测”任务数据集[1]。其余未公开数据集多以语文作文、古代诗歌、高考试卷、中文微博数据集等作为语料来源。现有的数据集语料规模小，修辞类别不均衡，限制了修辞识别任务的研究，因此建立更全面和更大规模的修辞数据集对于推动修辞识别任务的发展具有重要意义。中文修辞手法多样且概念差异大，为建立规模合理、类别平衡且修辞覆盖性广泛的修辞数据集，探究大语言模型对多种常见修辞手法的认知能力，本文采用人工标注的方式构建了一个多类别中文修辞句数据集MCRSD，构建流程如图2所示。未来，可以通过人工标注或者机器辅助生成的方式对数据集进行扩充，进一步提高该数据集的应用价值。

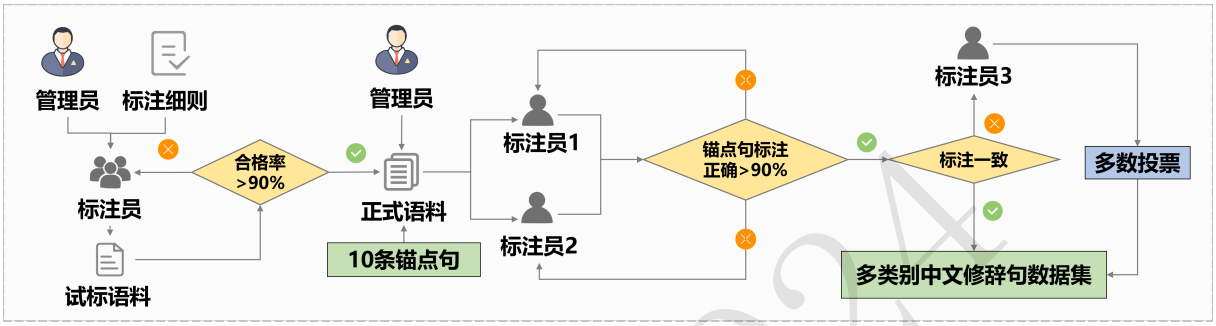



图 2: 多类别中文修辞句数据集构建流程

MCRSD数据集构建以北京语言大学BCC现代汉语语料库(荀恩东et al., 2016)、北京大学CCL语料库(詹卫东et al., 2019)和Li et al.(2022b)构建的CLGC文采优美段语料库作为基础，以句为单位进行切分作为待标数据。本次标注招募三名语言学专业的硕士研究生作为标注员，标注过程分为标注培训、预标注、正式标注、核对检验四个环节。在标注培训环节，标注员需要认真学习十种修辞的相关理论，明确标注任务和标注细则，标注过程中如果待标句子不具有任何修辞手法则标注为“无修辞”。培训完成后，每名标注员分配100条试标语料进行试标注，标注准确率达到90%方为合格。在正式标注阶段，管理员每天发放500条标注语料，两名标注员进行独立标注。为确保标注质量，在正式语料中投放10条锚点句，管理员对每日的标注数据进行整理检查，锚点句标注正确率达到90%以上且两名标注员标注一致的数据则纳入数据集。标注不一致的数据则发放给第三名标注员进行标注，采用多数投票原则选出句子最终标签，若无法选出则删除该条语料。持续以上标注过程，直至十种修辞和“无修辞”标签均具有1000条标注句子。最终，通过标注收集到了11000条标签句，MCRSD数据集的情况如表1所示。

| 类别 | 数量 | 平均词类符数 | 平均句长 | 平均实词数 | 平均虚词数 | 示例 |
|---|---|---|---|---|---|---|
| 比喻 | 1000 | 20.35 | 34.96 | 16.50 | 6.46 | 他沉默着，像一个木偶似的站立在林青史的面前。 |
| 比拟 | 1000 | 16.09 | 26.96 | 12.16 | 5.37 | 这会儿，你看，小草含着泪珠儿，在泣在愁。 |
| 排比 | 1000 | 22.17 | 48.51 | 25.75 | 7.62 | 夏天的夜晚，是那么的宁静，那么的美丽，那么凉爽。 |
| 夸张 | 1000 | 16.79 | 25.67 | 12.72 | 5.11 | 一个芝麻大的官放个屁，到了你们这儿也是8级地震！ |
| 反问 | 1000 | 14.50 | 22.92 | 10.76 | 4.60 | 下回有事谁还敢找你呀！ |
| 设问 | 1000 | 16.72 | 30.36 | 14.36 | 5.52 | 是谁创造了人类世界？是我们劳动群众。 |
| 对偶 | 1000 | 7.73 | 13.70 | 7.24 | 1.11 | 无尽波涛归学海，长春花木在词林。 |
| 反复 | 1000 | 12.92 | 23.15 | 12.16 | 5.10 | 冒着敌人的炮火，前进！前进！前进！ |
| 通感 | 1000 | 16.03 | 26.55 | 12.30 | 5.11 | 话语消融在拂过的轻风和升起的湿润香气之中。 |
| 顶真 | 1000 | 10.00 | 17.63 | 9.64 | 2.32 | 味摩诘之诗，诗中有画；观摩诘之画，画中有诗。 |
| 无修辞 | 1000 | 15.13 | 24.34 | 11.88 | 4.64 | 是这样，我回去替你想想办法吧，明天再告诉您，好吗? |

表 1: MCRSD数据集概况

[0]https://github.com/FlagOpen/FlagEval
[1]https://github.com/cubenlp/CERRU


(c) 2024 中国中文信息学会计算语言学专业委员会

表1展示了MCRSD数据集中每种标签类别的数量、平均词类符数、平均句长、平均实词数、平均虚词数和示例。通过分析这些统计数据，可以看出MCRSD数据集中不同类别的修辞句子在语言使用上的特点与差异。例如，排比句的平均词类符数（22.17）、平均句长（48.51）和平均实词数（25.75）相对较高，表明了排比修辞的句子结构相对其他修辞句更加复杂多样。而对偶句的各项指标数值都相对较低，体现了其简洁对仗的特点。比喻、比拟、夸张、设问和通感修辞句的平均词类符数都相对较高，分别为20.35、16.09、16.79、16.72和16.03，说明这些修辞手法倾向于使用复杂语言结构，增强了描述的生动性和多样性。

### 3.2 大模型修辞认知评测集组成与测试

针对MCRSD数据集中每条修辞标签句，随机从“无修辞”标签句中抽取3条句子与其组合成选择题，题目是“请从以下四个选项中选出具有XX修辞手法的句子”，候选答案包括ABCD四个选项，其中包括1条正确修辞标签句和3条“无修辞”标签句，组建形成了包含10000个修辞问题的评测集。评测集基本统计信息如附录A表7所示。选取ChatGPT4(gpt-4-1106-preview)(Achiam et al., 2023)、ChatGPT3.5(gpt-3.5-turbo)、文心一言(ERNIE-Bot-turbo-0922)[2]、Baichuan2-7B-Chat[3]、ChatGLM3-6B[4]、Qwen-7B-Chat[5]六种不同的大语言模型进行初步测试，使用准确率指标来评估大语言模型对修辞手法的认知能力。

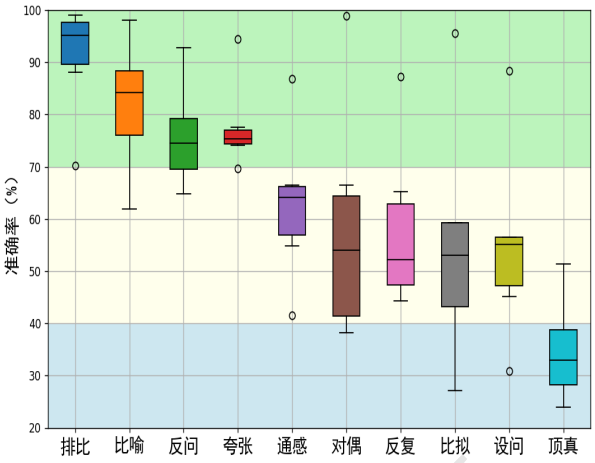



图 3: 不同模型在十种修辞上的平均识别效果

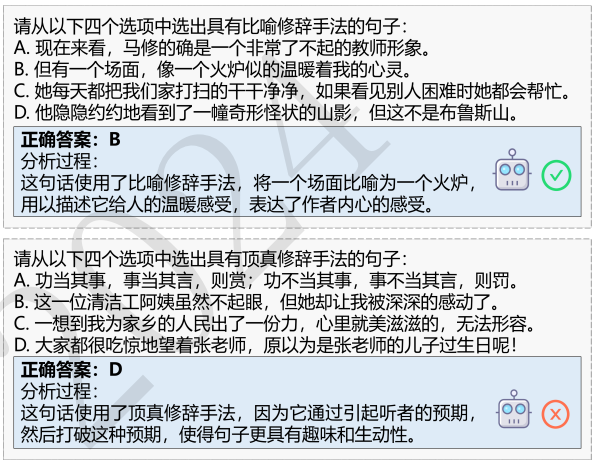



图 4: ChatGPT3.5评测示例

图3展示了六个大语言模型在十种修辞上的平均识别效果。可以看出，大语言模型对十种修辞的理解存在认知级差。对于排比、比喻等修辞，大语言模型呈现较高的认知水平，其识别准确率普遍达到70%以上；对于通感、对偶等修辞，大语言模型的修辞认知能力一般，识别准确率在70%至40%之间；而对于顶真修辞，大语言模型的认知水平相对较低，其识别准确率仅在40%以下。图4展示了ChatGPT3.5在测试过程中回答正确和回答错误的示例。可以发现，对于比喻，模型能够很好地分析出比喻的本体、喻体及语言效果；但对于顶真，模型不能准确理解该修辞手法的定义或特征，可能的原因是在模型的训练数据中某些修辞出现频率较低，导致模型受到数据偏差的影响，在这些修辞上的识别表现相对较差。这也说明了大语言模型的修辞认知能力和其修辞识别性能存在密切相关性，如果大语言模型对某种修辞具有较高的认知能力，那大语言模型对应的修辞识别性能也会更强。

### 3.3 多类别中文修辞知识库MCRKB构建

为确保后续为大语言模型补充领域信息的规范性和针对性，有效增强语言模型的修辞认知能力，本文参考陈望道(2008)、骆小所(2010)和谭汝为(2023)等人提出的修辞理论并结合多个知识来源构建了包括比喻、比拟、排比、夸张、设问、反问、对偶、反复、通感、顶真十种修辞手法的修辞知识库MCRKB，其中每类修辞的领域信息划分为定义、特征、变体特征、分析方

[2] https://yiyan.baidu.com/welcome
[3] https://www.baichuan-ai.com/home
[4] https://chatglm.cn/main/detail
[5] https://tongyi.aliyun.com/qianwen/


(c) 2024 中国中文信息学会计算语言学专业委员会

法、易混淆点、作用六个方面。MCRKB修辞知识库的基本统计信息如附录A表8所示，其中比喻修辞的知识点示例如附录A表9所示。MCRKB修辞知识库构建完成后，我们邀请两名语言学专业的硕士研究生进行人工审核和校对，确保知识库中信息的准确性、可靠性和权威性。图5展示了从MCRKB修辞知识库中提取相关修辞领域信息，增强大语言模型修辞认知能力的方式。对比图4中的模型回答错误示例发现，模型的修辞识别效果有所增强，验证了MCRKB修辞知识库的可用性。

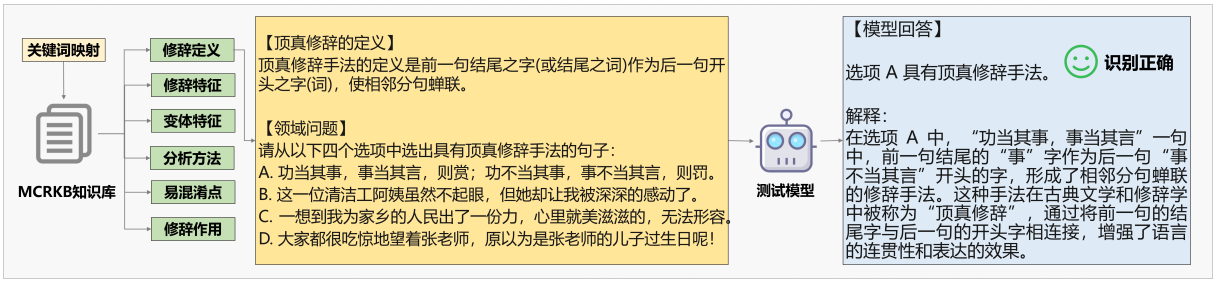



图 5: MCRKB修辞知识库为大语言模型补充领域信息的方式

## 4 方法

### 4.1 QAKAG整体框架

受限于数据集质量和模型性能，传统识别方法并不具备对修辞的理解认知能力。大语言模型通过在海量文本数据上的训练，具有较强的语言分析和语义理解能力，能够为解决复杂认知问题提供了新的可能性。在这样的背景下，本文提出了QAKAG框架，致力于系统化地识别并弥补大语言模型在特定领域知识理解方面存在的不足，通过领域信息认知缺陷检测与动态补充信息机制，增强大语言模型对领域知识的认知水平。整体框架如图6所示。

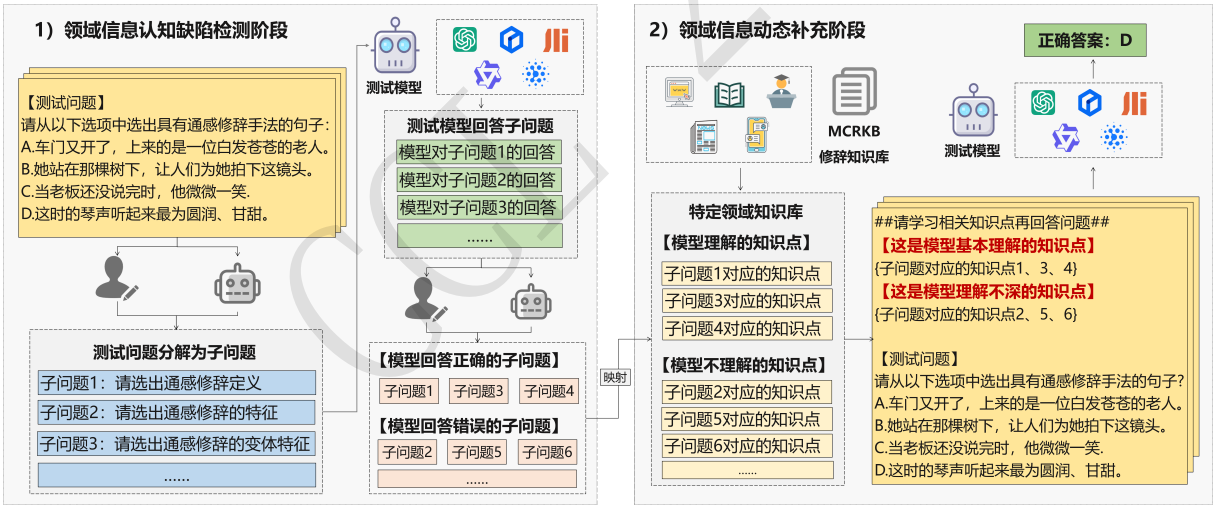



图 6: QAKAG整体框架

QAKAG框架分为：领域信息认知缺陷检测阶段和领域信息动态补充阶段。在领域信息认知缺陷检测阶段，将输入的测试问题分解为一系列子问题，判断大语言模型对每一个子问题的回答，获取大语言模型在领域知识上的理解情况。该阶段可以揭示大语言模型对哪些知识点已经基本了解，对哪些知识点需要加强学习，为后续的信息补充提供目标指向。然后，在领域信息动态补充阶段中，基于第一阶段获得的大语言模型对知识点理解情况，对大语言模型基本理解和理解不足的领域知识进行针对性的信息补充。本文所提出的QAKAG框架不仅可以强化大语言模型对已掌握知识点的理解，而且能够补充大语言模型在特定领域的知识缺陷。在QAKAG框架的增强下，大语言模型能够对不同修辞手法有更加深入的认知理解。

### 4.2 领域信息认知缺陷检测阶段

领域信息认知缺陷检测阶段的核心目的是评估大语言模型对子问题的认知能力。首先，通过“人工+模型”的方法，将测试问题分解为若干个具有高度针对性的子问题，并为子问题设置一项正确答案和三项错误答案。具体来说，本文通过设计子问题分解提示模板，使用ChatGPT4模型将测试问题从[定义、特征、变体特征、分析方法、易混淆点、作用]六个层面进行分解。由于大语言模型对修辞的理解深度不够，分解出来的子问题候选答案可能存在不准确的情况，因此在这一步加入人工校对步骤，对大语言模型分解出来的子问题进行审核。子问题若分解准确则进入下一步，子问题若有问题，则进入编辑模式修正该子问题，然后再进入下一步。以上阶段每类修辞仅需执行一次，审核通过的子问题将自动保存到子问题库中，后续若遇到相同类型的修辞问题，则自动从子问题库中进行抽取对应的子问题。这些子问题也分别对应MCRKB修辞知识库中六个方面的领域信息。测试模型被要求逐一回答这些子问题，以便判定其是否掌握回答测试问题所需的领域知识。子问题库形式如表2所示，每个子问题包括一项正确答案和三项错误答案，答案列是子问题的正确答案。

| 子问题 | 答案 |
|---|---|
| 请从以下选项中选出“顶真”修辞手法的定义:<br>A.把物当作人写，或把人当作物写，或把甲物当作乙物来写。<br>B.为了加强语气，用疑问的形式表示确定的意思，无疑而问，明知故问。<br>C.把三个或三个以上结构相同或相似的句子或句子成分排列起来。<br>D.前一句结尾之字(或结尾之词)作为后一句开头之字(词)，使相邻分句蝉联。 | D |

表 2: 子问题示例

测试模型对子问题的回答是否正确将被用作衡量其领域知识理解度的指标。如果测试模型能够正确回答子问题，可以推断测试模型已经掌握了该子问题对应的知识点，具备利用该知识点进行有效推理和应用的能力；反之，如果测试模型未能正确回答子问题，则推断测试模型尚未理解相关知识点或存在理解偏差，可能会造成测试模型在后续推理过程中的逻辑错误。通过这种分解和评估机制，我们能够检测出测试模型对解答测试问题所需领域知识的掌握水平，为进一步优化测试模型在领域知识上的理解和应用提供明确方向。

### 4.3 领域信息动态补充阶段

在给大语言模型输入领域信息时，不同的知识组合方式可能会对其输出产生不同的影响。本文设计了四种不同的知识组合方式，分别是全知直授法、断点补教法、诊断增强法和辩知强化法，如图7所示。对评测集中第$n$个问题，通过领域信息认知缺陷检测阶段分解为$m$个子问题。根据模型对子问题的回答判断其对知识点的理解情况，如果子问题回答正确则认为模型基本理解该知识点，反之则推断模型对该知识点理解不足。将子问题回答情况映射到知识库中，可以将$K$个相关知识点划分为$K_{i(\text{true})}$基本理解和$K_{i(\text{false})}$理解不足两种类别。

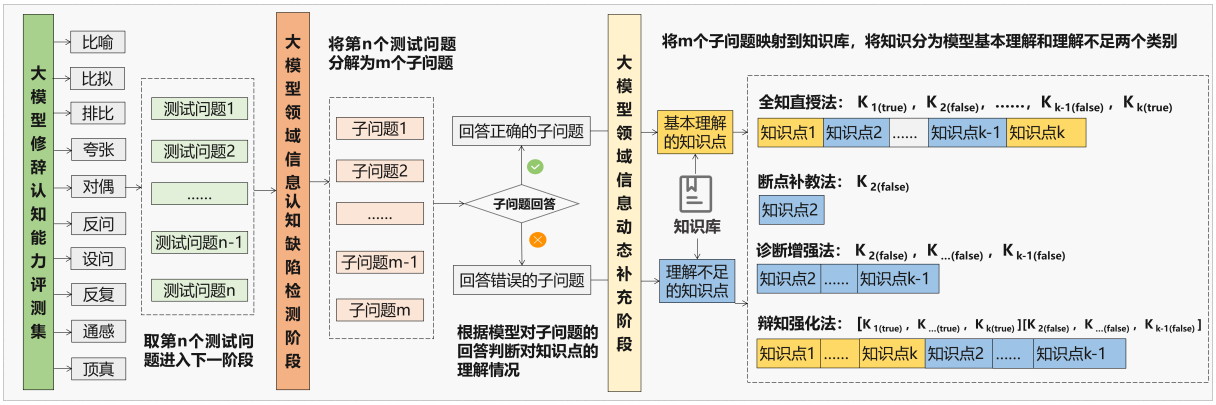



图 7: 在领域信息动态补充阶段设计四种知识组合方式

全知直授法，指不区分大语言模型对知识点的理解情况，一次性向大语言模型补充相关知识点。断点补教法，指仅补充大语言模型回答错误的第一个子问题所对应的知识点。诊断增强法，指补充大语言模型回答错误的所有子问题所对应的知识点。辩知强化法，指根据大语言模型对知识点的理解情况，将基本理解和理解不足的知识点分别进行组合，然后再与测试问题结合输入给大语言模型。通过尝试不同的知识组合方式，让大语言模型进行知识的理解和补全，有助于确定出最佳的知识组合方式，更好地提高模型的认知能力。

为确保新补充的知识能够被模型有效理解，本文利用以上设计的四种知识组合方式进行实验，最终选定辩知强化法作为QAKAG框架领域信息动态补充阶段的知识组合方式，补充形式如：$[K_{1(\text{true})}, K_{...(\text{true})}, K_{k(\text{true})}][K_{2(\text{false})}, K_{...(\text{false})}, K_{k-1(\text{false})}][n]$。这种方式将引导模型对不同知识点的关注方向，强化对基本理解知识点的掌握，并提高对理解不足知识点的认知。

领域信息动态补充阶段的目标在于设计有效的信息补充机制。根据认知缺陷检测阶段中模型对子问题的回答情况，以关键词映射的方式从知识库中提取出子问题对应的知识点，以此优化模型的知识结构并扩展其知识边界，帮助模型达到更高层次的认知和分析水平。关键词映射方式，即测试问题在分解为系列子问题后，子问题的题目形如“请从以下四个选项中选出比喻修辞的定义”。首先，检测该子问题的题目部分中是要求回答[比喻，比拟，夸张，排比，反复，反问，设问，对偶，通感，顶真]中的哪一修辞类别，检测出来的类别即作为该测试问题的关键词1。然后，继续检测该子问题的题目部分中要求回答[定义，特征，变体特征，分析方法，易混淆点，作用]中的哪一知识点类型，检测出来的类型即作为该测试问题的关键词2。最后，以关键1和关键2映射到MCRKB知识库中，即可提取到子问题对应的知识点。

## 5 实验与结果

### 5.1 实验设计

实验环节选用的测试模型包括ChatGPT4(gpt-4-1106-preview)、ChatGPT3.5(gpt-3.5-turbo)、文心一言(ERNIE-Bot-turbo-0922)、Baichuan2-7B-Chat、ChatGLM3-6B、Qwen-7B-Chat六个大语言模型。其中ChatGPT4、ChatGPT3.5、文心一言通过API调用的方式进行测试，Baichuan2-7B-Chat、ChatGLM3-6B、Qwen-7B-Chat则是从https://huggingface.co/网站下载开源代码，部署到本地服务器上进行测试。实验数据是包含10000个修辞问题的评测集。实验采用精确率（P）、召回率（R）和F1值作为评价指标。

**大语言模型修辞认知能力探究** 探究大语言模型在未经任何特定训练或提示优化的条件下在修辞问题上的认知表现。在不附加任何形式的指导信息下，将修辞识别问题直接输入大语言模型，探究大语言模型在原始状态下处理修辞识别问题的表现，为后续的优化提供基准，提示形式如下：[原始测试问题]→模型输出。

**提示策略增强大语言模型修辞认知能力探究** 选取提示工程策略中最具有代表性的零样本思维链（Zero-shot-CoT）、检索增强生成（RAG）、少样本学习（FSL）方法来评估大语言模型修辞认知能力增强效果。在大语言模型解决修辞识别问题时，Zero-shot-CoT方法是在测试问题结尾附加“让我们一步步思考(let’s think step by step)”，提示形式如下：[原始测试问题][Zero-shot-CoT]→模型输出。RAG方法是以关键词检索的方式，从百度百科[6]数据源中检索信息辅助大语言模型生成答案，提示形式如下：[百度百科信息提示][原始测试问题]→模型输出。FSL方法是以关键词映射的方式，从自建的修辞问题样本库中抽取5个示例样本融入提示中，增强大语言模型快速学习能力，提示形式如下：[示例1-5] [原始测试问题]→模型输出。

**QAKAG框架增强大语言模型修辞认知能力探究** 采用本文提出的QAKAG框架，在六个大语言模型上进行测试。通过QAKAG框架，探究大语言模型对测试问题分解出的系列子问题的理解情况，并将映射得到的领域信息组合后与原始测试问题合并输入大语言模型，以此验证QAKAG框架对大语言模型修辞认知能力的增强效果，该实验方法即为Ours。同时，采用全知直授法、断点补教法、诊断增强法进行对比实验。

### 5.2 实验结果及分析

ChatGPT等大语言模型在文本分类任务中具有巨大的潜力，可以有效解决流派识别、情感分析、立场检测等问题(Liu et al., 2023)，但尚未有研究工作关注如何利用大语言模型处理多类别中文修辞认知问题。因此，本文利用组建的评测集深入探索了不同大语言模型在不同修辞

[6] https://baike.baidu.com/


(c) 2024 中国中文信息学会计算语言学专业委员会

手法上的认知能力。表3展示了不同方法下大语言模型认知结果。从表中可以看出，对比其他方法，六个大语言模型均在Ours上取得了最优的性能，表明本文所提出的QAKAG框架对增强大语言模型的修辞认知能力具有更好的效果。其中，ChatGPT4在不同方法下都取得了最优异的表现。在直接回答的情况下，ChatGPT4的F1值就达到了87.5%，且在QAKAG框架的增强下其F1值提高到了95.8%。这说明ChatGPT4本身具有比较高的基础性能，在QAKAG框架下，能够对修辞进行更加全面、深入的认知理解。

| 分析方法 | ChatGPT4 | ChatGPT3.5 | 文心一言 | Baichuan2-7B-Chat | ChatGLM3-6B | Qwen-7B-Chat |
|---|---|---|---|---|---|---|
| | P/R/F1 | P/R/F1 | P/R/F1 | P/R/F1 | P/R/F1 | P/R/F1 |
| 直接回答 | .877/.875/.875 | .645/.612/.610 | .621/.612/.698 | .618/.532/.552 | .600/.589/.591 | .725/.686/.698 |
| Zero-shot-CoT | .894/.867/.879 | .672/.609/.637 | .583/.389/.361 | .636/.578/.592 | .684/.639/.658 | .774/.727/.737 |
| RAG-BaiKe | .920/.912/.915 | .753/.694/.722 | .628/.468/.460 | .705/.370/.334 | .697/.666/.664 | .628/.538/.564 |
| Few-Shot5 | .913/.912/.911 | .869/.868/.868 | .712/.332/.256 | .723/.575/.566 | .815/.796/.797 | .784/.778/.777 |
| 全知直授法 | .958/.956/.957 | .874/.870/.870 | .682/.610/.620 | .672/.621/.614 | .709/.681/.680 | .725/.717/.717 |
| 断点补教法 | .932/.929/.930 | .835/.820/.818 | .658/.575/.597 | .714/.631/.620 | .748/.687/.709 | .801/.784/.788 |
| 诊断增强法 | .957/.955/.956 | .885/.874/.874 | .660/.590/.602 | .741/.630/.630 | .709/.679/.685 | .718/.696/.702 |
| **Ours** | **.960/.957/.958** | **.888/.885/.885** | **.747/.743/.745** | **.737/.649/.646** | **.819/.801/.800** | **.844/.829/.829** |

表 3: 不同方法下大语言模型在修辞识别任务上的表现

从表3的实验结果中，对比提示策略可以发现，六个大语言模型在修辞识别任务上的整体认知表现：Few-Shot5>Zero-shot-CoT>RAG-BaiKe。相较直接回答，在Few-Shot5和Zero-shot-CoT提示策略下，除文心一言的F1值有所下降外，其余大语言模型的整体认知表现均有一定程度的提升，这可能是文心一言的学习推理能力稍有不足。而RAG-BaiKe提示策略对各大语言模型的增强效果不佳，则可能是因为从百度百科返回的领域信息长且冗余，缺少针对性和规范性，存在噪音问题，导致模型性能下降。对于不同的知识组合方式，对比全知直授法、断点补教法、诊断增强法三种不同的知识组合方式，采用辩知强化法的Ours可以帮助模型具有更好的识别性能，表明在解决复杂认知问题时，将知识点按照大语言模型基本理解和理解不足进行区分后再进行补充，有利于模型对知识点进行针对性学习，从而有更深入的理解。

从图8中可以看出，ChatGPT3.5在QAKAG框架下的F1值相对于直接回答增幅为45.1%，提升幅度是六个模型里面最显著的，说明ChatGPT3.5的训练数据中可能存在修辞知识不足或不平衡问题，但模型对上下文信息的敏感度和学习性能很高，QAKAG框架可以帮助模型更好地掌握并应用新的知识信息。值得注意的是，ChatGLM3-6B的增幅为35.4%，仅次于ChatGPT3.5，表明ChatGLM3-6B具有较强的学习推理能力，能快速分析领域信息并完成自我增强，在处理认知问题上具有很大的潜力。其余四个模型也都在QAKAG框架下有了不同幅度的提升，且对比其他提示策略和知识组合方式均达到了最优，验证了QAKAG框架的有效性。图9展示了ChatGPT3.5在采用QAKAG框架前后的回答样例，可以看出ChatGPT3.5通过QAKAG框架的增强，可以很好地学习到顶真修辞的相关知识，从而正确回答测试问题。

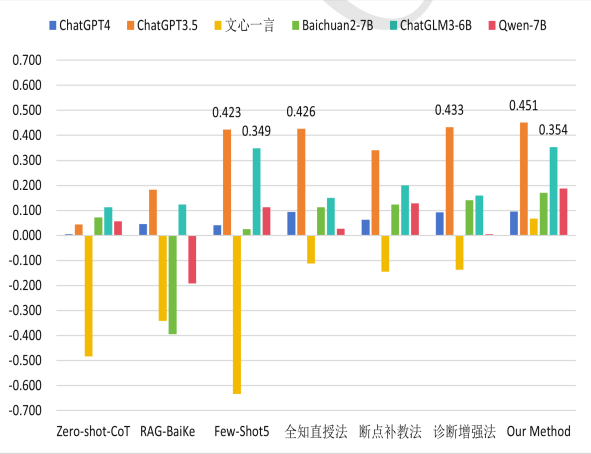



图 8: 各方法相对于直接回答的变化幅度对比

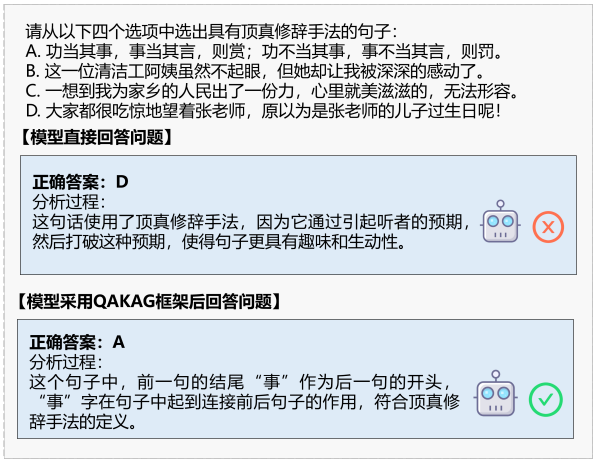



图 9: 模型应用QAKAG框架前后回答对比


(c) 2024 中国中文信息学会计算语言学专业委员会

### 5.3 探究实验分析

**(1) QAKAG框架各组成部分对大语言模型修辞认知能力的影响**

为深入探究QAKAG框架各个组成部分对实验结果的作用和影响，本文设计了三个消融实验进行验证。在第一个消融实验中，我们消除了领域信息动态补充阶段，即将大语言模型对子问题的回答情况补充在输入提示中辅助大语言模型生成答案，提示形式如下：[子问题][模型回答正误][子问题正确答案][原始测试问题]→模型输出，从而验证领域信息动态补充阶段的有效性。在第二个消融实验中，我们消除了领域信息缺陷检测阶段，即不通过子问题分解检测，直接以关键词映射的形式从知识库中将问题对应的知识点补充给大语言模型，提示形式如下：[测试问题相关知识点][原始测试问题]→模型输出，从而验证领域信息缺陷检测阶段的有效性。在第三个消融实验中，为验证QAKAG框架中MCRKB修辞知识库的有效性，我们将MCRKB修辞知识库分别替换为ChatGPT4知识库和百度百科知识库，比较不同来源的知识库对框架效果的影响。ChatGPT4知识库，即将第一阶段分解出的子问题输入给ChatGPT4，将模型生成的答案作为补充的领域信息来源；百度百科知识库，则是利用第一阶段分解出的子问题，搜索百科百科相关条目作为补充的领域信息来源，若百度百科没有对应的条目，则该子问题则跳过，不补充相关知识。表4展示了不同的消融实验下大语言模型在修辞识别任务上的认知表现结果。

| 分析方法 | ChatGPT4 | ChatGPT3.5 | 文心一言 | Baichuan2-7B-Chat | ChatGLM3-6B | Qwen-7B-Chat |
|---|---|---|---|---|---|---|
| | P/R/F1 | P/R/F1 | P/R/F1 | P/R/F1 | P/R/F1 | P/R/F1 |
| **Ours** | **.960/.957/.958** | **.888/.885/.885** | **.747/.743/.745** | .737/**.649/.646** | **.819/.801/.800** | **.844/.829/.829** |
| 消融领域信息动态补充 | .943/.939/.940 | .885/.880/.881 | .690/.483/.477 | .724/.593/.582 | .797/.768/.769 | .814/.807/.808 |
| 消融领域信息缺陷检测 | .958/.956/.957 | .874/.870/.870 | .682/.610/.620 | .672/.621/.614 | .709/.681/.680 | .725/.717/.717 |
| 替换MCRKB知识库-ChatGPT4 | .902/.900/.900 | .745/.741/.741 | .697/.505/.511 | .637/.464/.440 | .705/.687/.689 | .686/.641/.650 |
| 替换MCRKB知识库-BaiKe | .911/.910/.909 | .871/.863/.864 | .719/.715/.716 | **.759**/.497/.476 | .770/.742/.738 | .808/.791/.791 |

表 4: 各消融实验下大语言模型在修辞识别任务上的表现

可以发现，对比QAKAG框架的增强效果，如果消融领域信息动态补充阶段，六个大语言模型在修辞识别性能的平均F1值下降了6.77%，表明模型在某些修辞手法的理解上缺乏相关领域信息，对不同修辞手法存在认知不全或不平衡问题。如果消融领域信息缺陷检测，模型修辞识别性能的平均F1值下降了6.75%，这可能是大语言模型在学习相关知识时缺乏针对性，出现了一定程度的信息损失。如果将MCRKB知识库替换为ChatGPT4知识库，模型修辞识别性能的平均F1值则下降了15.60%，说明ChatGPT4对修辞知识的理解和掌握还不足够准确；如果将MCRKB知识库替换为百度百科知识库，模型修辞识别性能的平均F1值则下降了6.15%，说明百度百科对修辞知识的内容组织、质量或者覆盖范围上与MCRKB知识库存在一定的差距。故无论消融或替换任一个模块，各大语言模型的修辞识别能力均有所下降，验证了QAKAG框架中两个阶段的有效性和人工构建MCRKB修辞知识库的合理性。

**(2) QAKAG框架增强大语言模型对不同修辞手法认知能力的影响**

大语言模型对不同修辞手法的理解程度存在差异。表5和表6分别展示了Baichuan2-7B-Chat和文心一言在不同方法下对十种修辞手法的识别性能，括号内的数字代表相对于直接回答，模型性能的变化情况。可以看到，除了Ours外，其余方法下Baichuan2-7B-Chat和文心一言的修辞识别性能均有不同程度的提升和下降。附录A中的表10、11、12、13也展示了其他四个大语言模型的表现情况。综合所有数据发现，在QAKAG框架的增强下，六个大语言模型在十种修辞手法上的识别性能均有提升，表明了QAKAG框架的鲁棒性和普适性。

| 修辞类别 | Zero-shot-CoT | RAG-BaiKe | Few-Shot5 | 全知直授法 | 断点补教法 | 诊断增强法 | Ours |
|---|---|---|---|---|---|---|---|
| 排比 | .800(.098↑) | .763(.061↑) | .772(.070↑) | .887(.195↑) | .909(.217↑) | .673(.019↓) | .790(.098↑) |
| 比喻 | .759(.022↑) | .394(.343↓) | .717(.020↓) | .765(.028↑) | .761(.024↑) | .791(.054↑) | .787(.050↑) |
| 夸张 | .693(.064↓) | .499(.258↓) | .691(.066↓) | .626(.131↓) | .679(.078↓) | .696(.061↓) | .772(.015↑) |
| 反问 | .769(.018↑) | .513(.238↓) | .702(.049↓) | .802(.051↑) | .682(.069↓) | .740(.011↓) | .772(.021↑) |
| 比拟 | .499(.029↑) | .393(.077↓) | .673(.203↑) | .435(.035↓) | .633(.163↑) | .638(.168↑) | .653(.183↑) |
| 对偶 | .500(.118↑) | .261(.121↓) | .586(.204↑) | .631(.249↑) | .498(.116↑) | .673(.291↑) | .754(.372↑) |
| 通感 | .586(.171↑) | .634(.219↑) | .508(.093↑) | .559(.144↑) | .685(.270↑) | .673(.258↑) | .605(.190↑) |
| 反复 | .463(.007↓) | .345(.125↓) | .421(.049↓) | .513(.043↑) | .486(.016↑) | .603(.133↑) | .478(.008↑) |
| 设问 | .414(.105↑) | .477(.168↑) | .411(.102↑) | .431(.122↑) | .546(.237↑) | .400(.091↑) | .466(.157↑) |
| 顶真 | .296(.032↓) | .358(.030↑) | .273(.055↓) | .362(.034↑) | .433(.105↑) | .410(.082↑) | .408(.080↑) |

表 5: Baichuan2-7B-Chat在不同方法下对十种修辞手法的认知表现


(c) 2024 中国中文信息学会计算语言学专业委员会

| | Zero-shot-CoT | RAG-BaiKe | Few-Shot5 | 全知直授法 | 断点补教法 | 诊断增强法 | Ours |
|---|---|---|---|---|---|---|---|
| 排比 | .856(.025↓) | .830(.051↓) | .809(.072↓) | .856(.025↓) | .830(.051↓) | .809(.072↓) | .953(.072↑) |
| 比喻 | .655(.036↑) | .554(.065↓) | .636(.017↑) | .655(.036↑) | .554(.065↓) | .636(.017↑) | .858(.239↑) |
| 夸张 | .757(.060↑) | .703(.006↑) | .718(.021↑) | .757(.060↑) | .703(.006↑) | .718(.021↑) | .753(.056↑) |
| 反问 | .676(.005↓) | .683(.002↑) | .680(.001↓) | .676(.005↓) | .683(.002↑) | .680(.001↓) | .852(.171↑) |
| 比拟 | .640(.047↑) | .537(.056↓) | .601(.008↑) | .640(.047↑) | .537(.056↓) | .601(.008↑) | .668(.075↑) |
| 对偶 | .328(.252↓) | .364(.216↓) | .402(.178↓) | .328(.252↓) | .364(.216↓) | .402(.178↓) | .694(.114↑) |
| 通感 | .626(.078↑) | .589(.041↑) | .545(.003↓) | .626(.078↑) | .589(.041↑) | .545(.003↓) | .636(.088↑) |
| 反复 | .445(.002↑) | .406(.037↓) | .438(.005↓) | .445(.002↑) | .406(.037↓) | .438(.005↓) | .747(.304↑) |
| 设问 | .612(.048↑) | .600(.036↑) | .620(.056↑) | .612(.048↑) | .600(.036↑) | .620(.056↑) | .732(.168↑) |
| 顶真 | .501(.013↓) | .483(.031↓) | .450(.064↓) | .501(.013↓) | .483(.031↓) | .450(.064↓) | .541(.027↑) |

表 6: 文心一言(ERNIE-Bot-turbo-0922)在不同方法下对十种修辞手法的认知表现

图10展示了QAKAG框架对大语言模型在不同修辞手法认知上的增强效果。可以看出，测试的大语言模型对比喻、对偶等直观形象的修辞手法认知提升较为明显，而对于通感、反复等更依赖语境和逻辑关系的修辞手法的认知虽有所提升，但幅度略低。这可能是因为评测集采用了以句为单位的选择题形式，而对于逻辑丰富的修辞手法，大语言模型需要更多的上下文语境来增强认知。因此，对于大语言模型来说，要在这种片段化的评测中完全把握修辞手法的细微差别可能会更加困难。此外，实验结果也表明，通过QAKAG框架增强大语言模型修辞认知能力，大语言模型对不同修辞手法的理解和应用能力也会相应增强，其修辞识别性能也会同步提升，验证了大语言模型修辞认知能力和修辞识别性能之间的正相关性。

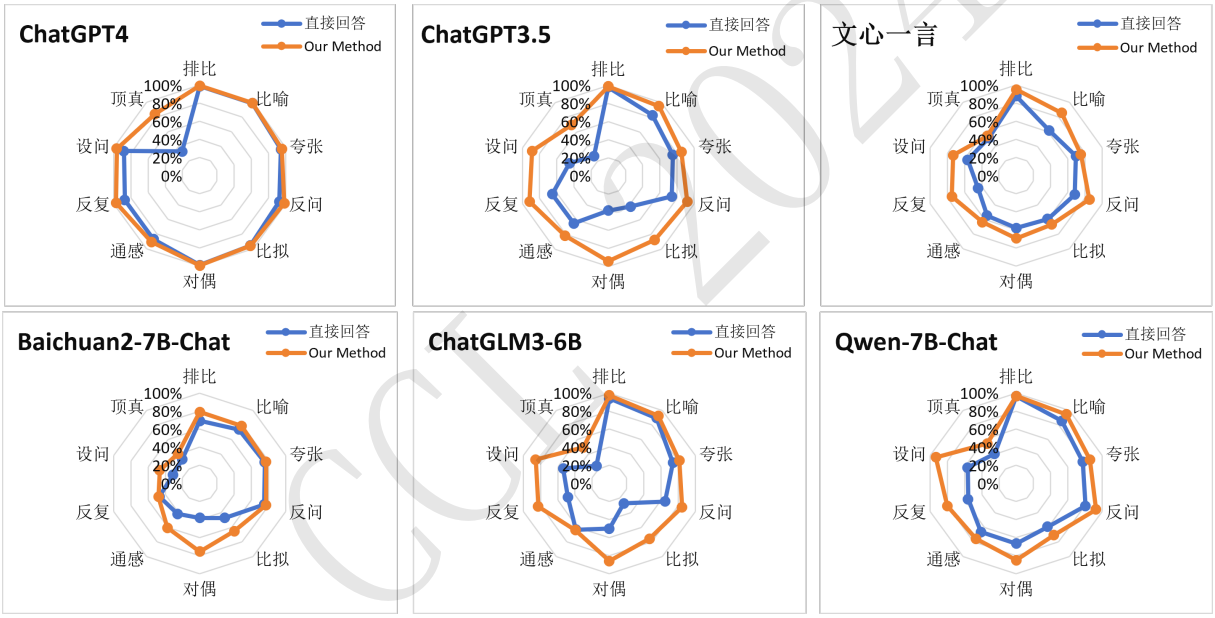



图 10: QAKAG框架对大语言模型在不同修辞手法识别能力增强情况

## 6 结论

针对大语言模型对不同修辞手法存在的认知级差问题，本文提出了QAKAG框架，提高大语言模型对不同修辞手法的认知能力，缩小其在不同修辞手法上呈现的认知差异，提升其修辞识别性能。为验证QAKAG框架的合理性，本文构建了多类别中文修辞句数据集MCRSD和修辞知识库MCRKB，并设计了四种不同的知识组合方式进行实验对比。实验证明，相比于直接回答和Zero-shot-CoT、RAG-BaiKe、Few-Shot5提示策略，QAKAG框架对大语言模型的修辞认知增强效果最优，而且各大语言模型对不同修辞手法的认知能力也均有不同程度的提升，表明QAKAG框架有效性的同时，也说明了大语言模型在处理复杂认知问题上具有强大潜力。但QAKAG框架也存在改进的空间，如减少问题分解阶段的人工介入、提高因提示信息的增加导致模型推理速度变慢的问题。未来，我们将以此为研究重点，持续优化QAKAG框架，提高其自动化程度和推理速度，并探索其在更广泛的修辞类别和语言理解任务中的适用性。


(c) 2024 中国中文信息学会计算语言学专业委员会

## 参考文献

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. "Gpt-4 technical report."arXiv preprint arXiv:2303.08774(2023).

Yigal Attali and Jill Burstein. "Automated essay scoring with e-rater® V. 2." The Journal of Technology, Learning and Assessment 4.3 (2006).

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, Furu Wei. "Why can gpt learn in-context? language models implicitly perform gradient descent as meta-optimizers." arXiv preprint arXiv:2212.10559 (2022).

Thomas K. Landauer, Darrel Laham, Peter W. Foltz. "Automated scoring and annotation of essays with the Intelligent Essay Assessor." Automated essay scoring: A cross-disciplinary perspective (2003): 87-112.

Yucheng Li, Chenghua Lin, and Frank Guerin. "Cm-gen: A neural framework for chinese metaphor generation with explicit context modelling." Proceedings of the 29th international conference on computational linguistics. 2022.

Yi Li, Dong Yu, and Pengyuan Liu. "CLGC: A corpus for Chinese literary grace evaluation." Proceedings of the Thirteenth Language Resources and Evaluation Conference. 2022.

Jinpeng Li, Zekai Zhang, Quan Tu, Xin Cheng, Dongyan Zhao, Rui Yan. "StyleChat: Learning Recitation-Augmented Memory in LLMs for Stylized Dialogue Generation." arXiv preprint arXiv:2403.11439 (2024).

Yangning Li, Shirong Ma, Xiaobin Wang, Shen Huang, Chengyue Jiang, Hai-Tao Zheng, Pengjun Xie, Fei Huang, Yong Jiang. "Ecomgpt: Instruction-tuning large language models with chain-of-task tasks for e-commerce." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 38. No. 17. 2024.

Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, Zihao Wu, et al. "Summary of chatgpt-related research and perspective towards the future of large language models." Meta-Radiology (2023): 100017.

Yanming Liu, Xinyue Peng, Xuhong Zhang, Weihao Liu, Jianwei Yin, Jiannan Cao, Tianyu Du. "RA-ISF: Learning to Answer and Understand from Retrieval Augmentation via Iterative Self-Feedback." arXiv preprint arXiv:2403.06840 (2024).

William C. Mann and Sandra A. Thompson. "Rhetorical structure theory: Toward a functional theory of text organization." Text-interdisciplinary Journal for the Study of Discourse 8.3 (1988): 243-281.

Ellis B. Page. "The imminence of... grading essays by computer." The Phi Delta Kappan 47.5 (1966): 238-243.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, Chris Dyer. "Metaphor detection with cross-lingual model transfer." Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2014.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, Ashish Sabharwal. "Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions." arXiv preprint arXiv:2212.10509 (2022).

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, Shuming Shi. "Siren's song in the AI ocean: a survey on hallucination in large language models." arXiv preprint arXiv:2309.01219 (2023).

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, Ed Chi. "Least-to-most prompting enables complex reasoning in large language models." arXiv preprint arXiv:2205.10625 (2022).

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, Diyi Yang. "Can large language models transform computational social science?." Computational Linguistics 50.1 (2024): 237-291.

蔡越著.AI赋能写作AI大模型高效写作一本通[M].北京：北京大学出版社, 2024


(c) 2024 中国中文信息学会计算语言学专业委员会

陈望道著.修辞学发凡[M].上海：复旦大学出版社, 2008

崔亿萍.大模型在自然语言处理的应用和研究[J].中国科技期刊数据库工业A, 2023, (12): 57-61

付瑞吉, 王栋, 王士进, 胡国平, 刘挺.面向作文自动评分的优美句识别[J].中文信息学报, 2018, 第32卷(6): 88-97

呼啸, 付瑞吉, 宋巍, 王士进, 胡国平, 秦兵, 刘挺.一种比喻句识别方法、装置、设备及存储介质[P].中国专利：CN111914544A, 2020.11.10

李玲玲, 王晓燕.人工智能翻译对文学文本处理的对比分析[J].时代报告, 2023, (46): 10-12

李明峰, 贾修一.基于多分类器集成学习的中文反语识别技术[J].计算机与数字工程, 2018, 第46卷(9): 1790-1795

刘广辉.基于修辞句识别的中学语文作文评价系统的设计与实现[D].中国科学院大学(中国科学院沈阳计算技术研究所), 2022

刘明杨.高考作文自动评分关键技术研究[D].哈尔滨工业大学, 2015

骆小所著.现代修辞学[M].昆明：云南人民出版社, 2010

邱文生.修辞的认知性研究[J].重庆理工大学学报(社会科学), 2015, (12): 122-126

石昀东.基于修辞使用的小学作文自动分类评价方法研究[D].华中师范大学, 2019

谭汝为著.汉语修辞学指要[M].天津出版传媒集团；天津人民出版社, 2023

王荣文著.现代汉语中的修辞艺术剖析[M].中国原子能出版社, 2018

徐盛桓.修辞研究的认知视角[J].西安外国语大学学报,2008,第16卷(2): 1-5

荀恩东, 饶高琦, 肖晓悦, 臧娇娇.大数据背景下BCC语料库的研制[J].语料库语言学, 2016, (1): 93-109, 118

詹卫东, 郭锐, 常宝宝, 谌贻荣, 陈龙.北京大学CCL语料库的研制[J].语料库语言学, 2019, (1): 71-86, 116

赵琳玲.面向高考鉴赏类问题的明喻与拟人识别方法研究[D].山西大学, 2020

朱晓亮, 谯宇同, 龚颖, 殷姿, 董鑫, 石昀东, 刘三女牙, 杨宗凯.一种小学语文作文排比句自动识别的方法[P].中国专利：CN110598202A, 2019.12.20

CCL 2024

# A 附录

| 标签 | 数量 | 平均长度 | 答案 | 问题示例 |
|---|---|---|---|---|
| 夸张 | 1000 | 144.952 | A | 请从以下四个选项中选出具有夸张修辞手法的句子：A. 江浩被她的惨叫吓得三魂冲天，七魂出窍。B. 卢老师告诉了我们一个小诀窍就是，先读一句然后背一句。C. 其实我觉得要留待下一期再和大家分享一下。D. 在我看来，马修是一个才华横溢的音乐家，不过在1949年的法国乡村，他没有发展自己才华的机会，最终成为了一间男子寄宿学校的助理教师。 |
| 通感 | 1000 | 145.081 | C | 请从以下四个选项中选出具有通感修辞手法的句子：A. 差5分零点时，庙前已聚集了数百人，他们在等待着那新旧交替的难忘一刻。B. 虽然我是她的得力助手，但她做事永远都是毫不留情的啊。C. 她的声音很透明。D. 我一怔，抬头正想回答，看见妈妈只穿了一件衣服，身子在寒风中瑟瑟发抖。 |
| 比拟 | 1000 | 144.095 | C | 请从以下四个选项中选出具有比拟修辞手法的句子：A. 我们班的李老师就是其中一位，她教语文，她对自己的每一节课都认真备课，仔细研究如何才能让学生更容易理解自己所讲的知识。B. 我很小的时候爸爸曾经给我讲过圣诞节的故事，那可真是一个美妙的故事啊！C. 落日挥毫泼墨，把我们的世界装扮得美丽动人。D. 人的劳动能力是生产力中最活跃、最重要的因素。 |
| 反复 | 1000 | 145.201 | B | 请从以下四个选项中选出具有反复修辞手法的句子：A. 我们在离开的时候，爸爸去了公平秤，称了那条鱼的重量，公平秤显示的是3斤。B. 英国人永远，永远，永远不会成为奴隶。C. 陈春泉答应了表妹王采玉的请求后，左思右想、反复比较，才选定了毛鼎和家。D. 前者倾向于突破现有规则，后者更愿意维护传统。 |
| 对偶 | 1000 | 130.715 | B | 请从以下四个选项中选出具有对偶修辞手法的句子：A. 虽然我是她的得力助手，但她做事永远都是毫不留情的啊。B. 把问题找实，把根源挖深。C. 我忍不住咬了几口，可它的汁液太少了，一颗不过瘾，又拿起第二颗，第三颗.....D. 我上前地问了一句：“需要我帮忙吗？” |
| 设问 | 1000 | 150.901 | C | 请从以下四个选项中选出具有设问修辞手法的句子：A. 我没有回答，我又想起了妈妈的话：将梦想进行到底。B. 我踏着轻松的步子来到教室里，同学们高高兴兴地纷纷跑上前来送我生日礼物。C. 女人同样喜欢帅哥，但又有多少男人愿意为了女人的一句赞美而天天待健身房，天天苦练肌肉？说到底这还是男女性社会地位问题不平等导致的。D. 然后，他又告诫自己，要冷静，不能感情用事。 |
| 比喻 | 1000 | 153.126 | B | 请从以下四个选项中选出具有比喻修辞手法的句子：A. 这时天空下起了绵绵细雨，但是人们一点也不在乎。B. 你就像辛勤的园丁，细心培育着祖国花朵，使这些花朵茁壮成长。C. 我生气极了，“不，你骗人，那不是故事，是真实的。”D. 明明是你暗恋别人，那个人从来察觉，也许知道了，却无动于衷。 |
| 反问 | 1000 | 140.766 | B | 请从以下四个选项中选出具有反问修辞手法的句子：A. 我爸爸最大的特点是：让我去挑战一些我不敢做的事情。B. 若非亲眼目睹，谁能相信水能变成“石头”呢？C. 就连她自己也说：“画动物是我的强项呦！”D. 从小到大可从来没有人送过礼物给我呀！ |
| 顶真 | 1000 | 135.800 | D | 请从以下四个选项中选出具有顶真修辞手法的句子：A. 作为马上要小学毕业的我，我们要好好珍惜我们之间的友谊，不要动不动就吵架。B. 每当看到大哥哥、大姐姐上学时那快乐的神情，我就羡慕极了。C. 大家都很吃惊地望着张老师，原以为是张老师的儿子过生日呢！D. 雪洁林，林护雪，雪影静谧。 |
| 排比 | 1000 | 173.492 | D | 请从以下四个选项中选出具有排比修辞手法的句子：A. 没有啊，我以为你喂了。B. 张老师高兴地让全班齐唱《祝你生日快乐》的歌。C. 这事情虽然看起来不可思议，但你们真的是冤枉潘武了。D. 延安的歌声它是黑夜的火把，雪天的煤炭，大旱的甘霖。 |

表 7: 大模型修辞认知评测集概况

| 修辞格 | 平均长度 | 最大长度 | 最小长度 | 知识点类型 |
|---|---|---|---|---|
| 反复 | 97.500000 | 192 | 41 | 定义、特征、变体特征、分析方法、易混淆点、作用 |
| 反问 | 95.833333 | 109 | 73 | |
| 夸张 | 109.500000 | 176 | 39 | |
| 对偶 | 94.666667 | 143 | 48 | |
| 排比 | 201.166667 | 532 | 55 | |
| 比喻 | 152.666667 | 228 | 48 | |
| 比拟 | 120.333333 | 183 | 66 | |
| 设问 | 94.333333 | 184 | 56 | |
| 通感 | 83.666667 | 96 | 57 | |
| 顶真 | 110.500000 | 288 | 45 | |

表 8: MCRKB知识库概况


(c) 2024 中国中文信息学会计算语言学专业委员会

| 修辞 | 类型 | 知识点 |
|---|---|---|
| 比喻 | 定义 | 比喻修辞手法的定义是根据心理联想，抓住和利用不同事物的相似点，用另一个事物来描绘所要表现的事物。 |
| 比喻 | 特征 | 比喻修辞手法的特征包括：1.一般由三部分组成，即：本体(被比喻的事物或情境)、喻词(表示比喻关系的词语)、喻体(比喻的事物或情境)。2.本体和喻体必须是性质不同的两类事物。3.在比喻句中，喻体必须出现。4.本体和喻体之间必须有相似点。比喻通过将不同的事物或情境进行类比，从而传达出相似之处，使得文学作品更加生动形象。 |
| 比喻 | 变体特征 | 比喻修辞手法的变体主要有三种：明喻、借喻、暗喻。明喻要求本体和喻体同时出现，但连接词为比喻词如“像”、“好似”、“仿佛”等，有时还会有配合词如“似的”、“一样”；暗喻要求本体和喻体同时出现，但使用了“是”、“成了”、“变成”等词来连接，或者以偏正式词组的形式出现，用本体修饰限制喻体，有时还会使用复指成分来构成暗喻；借喻则是直接叙述喻体，本体和喻词都不出现。 |
| 比喻 | 分析方法 | 比喻修辞手法的分析方法如下：例1：思想的春风吹遍油田。通过将“思想”与“春风”相比较，揭示出了它们之间的相似性，从而达到比喻的效果。例2：岸边的华灯倒映在湖中，宛如颗颗宝石缀在湖面之上。通过将“华灯”与“宝石”相比较，也实现了比喻的目的。例3：我就知道，我们之间已经隔了一层可悲的厚障壁了，我再也说不出话。通过描述“厚障壁”来暗示人们之间的隔阂，虽然没有直接提及喻词，但通过暗示和隐喻的方式也实现了比喻的效果。 |
| 比喻 | 易混淆点 | 比喻修辞手法易和比拟、夸张、通感修辞手法混淆。1.比喻与比拟之间的区别：比喻强调的是两个事物的相似性，而比拟则是利用它们的不同特性使它们融为一体，这一点在比喻中喻体是必不可少的，而在比拟中必须出现拟体。2.比喻与夸张之间的区别：夸张是放大或缩小事物的形象特征以增强表达效果，而比喻则是根据相似点将某一事物比作另一事物，使抽象的事物具体化，道理更加浅显易懂。3.比喻与通感之间的区别：比喻强调相似性，离不开本体、喻体和相似点，而通感更注重感觉器官之间的转移。 |
| 比喻 | 作用 | 比喻修辞手法的作用是通过将一个事物与另一个不同但相关的事物相比较，使深奥的道理更加浅显易懂，帮助人们加深体味；将抽象的事物具体化，使人们更容易接受；将概括的东西形象化，给人们留下鲜明的印象。 |

表 9: MCRKB知识库中比喻修辞的知识点示例

| | Zero-shot-CoT | RAG-BaiKe | Few-Shot5 | 全知直授法 | 断点补教法 | 诊断增强法 | Ours |
|---|---|---|---|---|---|---|---|
| 排比 | .995(.005↑) | .994(.004↑) | .995(.005↑) | .995(.005↑) | .994(.004↑) | .995(.005↑) | .995(.005↑) |
| 比喻 | .996(.006↑) | .987(.003↓) | .994(.004↑) | .996(.006↑) | .987(.003↓) | .994(.004↑) | .993(.003↑) |
| 夸张 | .955(.010↑) | .959(.014↑) | .959(.014↑) | .955(.010↑) | .959(.014↑) | .959(.014↑) | .957(.012↑) |
| 反问 | .987(.059↑) | .968(.040↑) | .982(.054↑) | .987(.059↑) | .968(.040↑) | .982(.054↑) | .984(.056↑) |
| 比拟 | .960(.005↑) | .940(.015↓) | .956(.001↑) | .960(.005↑) | .940(.015↓) | .956(.001↑) | .957(.002↑) |
| 对偶 | .995(.006↑) | .998(.009↑) | .997(.008↑) | .995(.006↑) | .998(.009↑) | .997(.008↑) | .995(.006↑) |
| 通感 | .915(.047↑) | .921(.053↑) | .916(.048↑) | .915(.047↑) | .921(.053↑) | .916(.048↑) | .908(.040↑) |
| 反复 | .970(.098↑) | .965(.093↑) | .957(.085↑) | .970(.098↑) | .965(.093↑) | .957(.085↑) | .974(.102↑) |
| 设问 | .964(.081↑) | .965(.082↑) | .964(.081↑) | .964(.081↑) | .965(.082↑) | .964(.081↑) | .965(.082↑) |
| 顶真 | .825(.493↑) | .590(.258↑) | .833(.501↑) | .825(.493↑) | .590(.258↑) | .833(.501↑) | .842(.510↑) |

表 10: ChatGPT4(gpt-4-1106-preview)在不同方法下对十种修辞手法的认知表现

| | Zero-shot-CoT | RAG-BaiKe | Few-Shot5 | 全知直授法 | 断点补教法 | 诊断增强法 | Ours |
|---|---|---|---|---|---|---|---|
| 排比 | .989(.009↑) | **.979(.001↓)** | .990(.010↑) | .989(.009↑) | **.979(.001↓)** | .990(.010↑) | .989(.009↑) |
| 比喻 | .945(.117↑) | .896(.068↑) | .922(.094↑) | .945(.117↑) | .896(.068↑) | .922(.094↑) | .953(.125↑) |
| 夸张 | .847(.096↑) | .808(.057↑) | .818(.067↑) | .847(.096↑) | .808(.057↑) | .818(.067↑) | .852(.101↑) |
| 反问 | .890(.151↑) | .885(.146↑) | .950(.211↑) | .890(.151↑) | .885(.146↑) | .950(.211↑) | .920(.181↑) |
| 比拟 | .835(.416↑) | .821(.402↑) | .773(.354↑) | .835(.416↑) | .821(.402↑) | .773(.354↑) | .874(.455↑) |
| 对偶 | .943(.558↑) | .859(.474↑) | .966(.581↑) | .943(.558↑) | .859(.474↑) | .966(.581↑) | .946(.561↑) |
| 通感 | .805(.154↑) | .812(.161↑) | .765(.114↑) | .805(.154↑) | .812(.161↑) | .765(.114↑) | .817(.166↑) |
| 反复 | .915(.263↑) | .865(.213↑) | .917(.265↑) | .915(.263↑) | .865(.213↑) | .917(.265↑) | .917(.265↑) |
| 设问 | .858(.407↑) | .770(.319↑) | .887(.436↑) | .858(.407↑) | .770(.319↑) | .887(.436↑) | .887(.436↑) |
| 顶真 | .676(.408↑) | .501(.233↑) | .754(.486↑) | .676(.408↑) | .501(.233↑) | .754(.486↑) | .695(.427↑) |

表 11: ChatGPT3.5(gpt-3.5-turbo)在不同方法下对十种修辞手法的认知表现


(c) 2024 中国中文信息学会计算语言学专业委员会

| | Zero-shot-CoT | RAG-BaiKe | Few-Shot5 | 全知直授法 | 断点补教法 | 诊断增强法 | Ours |
|---|---|---|---|---|---|---|---|
| 排比 | .975(.036↑) | .974(.035↑) | .966(.027↑) | .975(.036↑) | .974(.035↑) | .966(.027↑) | .974(.035↑) |
| 比喻 | .920(.028↑) | .880(.012↓) | .950(.058↑) | .920(.028↑) | .880(.012↓) | .950(.058↑) | .923(.031↑) |
| 夸张 | .840(.099↑) | .795(.054↑) | .788(.047↑) | .840(.099↑) | .795(.054↑) | .788(.047↑) | .815(.074↑) |
| 反问 | .885(.237↑) | .815(.167↑) | .763(.115↑) | .885(.237↑) | .815(.167↑) | .763(.115↑) | .847(.199↑) |
| 比拟 | .648(.376↑) | .549(.277↑) | .437(.165↑) | .648(.376↑) | .549(.277↑) | .437(.165↑) | .756(.484↑) |
| 对偶 | .903(.402↑) | .567(.066↑) | .725(.224↑) | .903(.402↑) | .567(.066↑) | .725(.224↑) | .859(.358↑) |
| 通感 | .644(.012↑) | .670(.038↑) | .629(.003↓) | .644(.012↑) | .670(.038↑) | .629(.003↓) | .639(.007↑) |
| 反复 | .820(.336↑) | .650(.166↑) | .651(.167↑) | .820(.336↑) | .650(.166↑) | .651(.167↑) | .828(.344↑) |
| 设问 | .871(.334↑) | .669(.132↑) | .608(.071↑) | .871(.334↑) | .669(.132↑) | .608(.071↑) | .857(.320↑) |
| 顶真 | .453(.214↑) | .296(.057↑) | .272(.033↑) | .453(.214↑) | .296(.057↑) | .272(.033↑) | .492(.253↑) |

表 12: ChatGLM3-6B在不同方法下对十种修辞手法的认知表现

| | Zero-shot-CoT | RAG-BaiKe | Few-Shot5 | 全知直授法 | 断点补教法 | 诊断增强法 | Ours |
|---|---|---|---|---|---|---|---|
| 排比 | .941(.024↓) | .978(.013↑) | .891(.074↓) | .941(.024↓) | .978(.013↑) | .891(.074↓) | .967(.002↑) |
| 比喻 | .943(.086↑) | .933(.076↑) | .782(.075↓) | .943(.086↑) | .933(.076↑) | .782(.075↓) | .945(.088↑) |
| 夸张 | .822(.047↑) | .845(.070↑) | .815(.040↑) | .822(.047↑) | .845(.070↑) | .815(.040↑) | .857(.082↑) |
| 反问 | .935(.129↑) | .867(.061↑) | .765(.041↓) | .935(.129↑) | .867(.061↑) | .765(.041↓) | .929(.123↑) |
| 比拟 | .744(.152↑) | .712(.120↑) | .644(.052↑) | .744(.152↑) | .712(.120↑) | .644(.052↑) | .708(.116↑) |
| 对偶 | .893(.228↑) | .788(.123↑) | .750(.085↑) | .893(.228↑) | .788(.123↑) | .750(.085↑) | .848(.183↑) |
| 通感 | .713(.048↑) | .747(.082↑) | .693(.028↑) | .713(.048↑) | .747(.082↑) | .693(.028↑) | .757(.092↑) |
| 反复 | .757(.197↑) | .691(.131↑) | .605(.045↑) | .757(.197↑) | .691(.131↑) | .605(.045↑) | .803(.243↑) |
| 设问 | .884(.319↑) | .747(.182↑) | .568(.003↑) | .884(.319↑) | .747(.182↑) | .568(.003↑) | .935(.370↑) |
| 顶真 | .490(.084↑) | .529(.123↑) | .446(.040↑) | .490(.084↑) | .529(.123↑) | .446(.040↑) | .545(.139↑) |

表 13: Qwen-7B-Chat在不同方法下对十种修辞手法的认知表现

CCL 2024

# 基于中间层对齐的异构师生模型知识蒸馏

**翟飞燕，王任之，李丕绩***
南京航空航天大学计算机科学与技术学院
模式分析与机器智能工业和信息化部重点实验室（南京航空航天大学）
江苏，南京，211106
{fyzhai, rzhwang, pjli}@nuaa.edu.cn

## 摘要

知识蒸馏技术作为大语言模型时代的一项前沿模型压缩策略，通过将复杂模型的知识有效迁移至简单模型，显著降低了模型的参数规模和计算成本。尽管如此，目前主流的生成式大语言模型蒸馏算法主要集中于优化师生模型间的最后输出层损失，而忽视了对模型中间层的探索。此外，针对中间层蒸馏的研究往往对师生模型的结构一致性有着严格的要求，无法处理异构模型间的蒸馏问题，从而存在明显的局限性。针对这些问题，我们提出了一种新的知识蒸馏算法：引入了中间层蒸馏损失的异构生成式师生大语言模型知识蒸馏算法。该算法首先提取师生模型的中间层信息作为蒸馏对象，随后通过专门设计的中间层映射规则和对齐模块，实现异构模型间基于中间层的知识对齐与损失计算。最后，联合优化各个蒸馏损失的比例。通过在五个相关数据集上进行实验验证，我们的方法在提高蒸馏效果方面展现出显著优势。

**关键词：** 生成式大语言模型；白盒知识蒸馏；中间层；异构性

# Knowledge distillation of heterogeneous teacher-student model with intermediate layer loss

**Feiyan Zhai, Renzhi Wang, Piji Li***
College of Computer Science and Technology,
Nanjing University of Aeronautics and Astronautics
MIIT Key Laboratory of Pattern Analysis and Machine Intelligence
Nanjing, 211106, China
{fyzhai, rzhwang, pjli}@nuaa.edu.cn

## Abstract

As a cutting-edge model compression strategy in the era of large language models, knowledge distillation technology significantly reduces the parameter scale and computational cost of models by effectively transferring the knowledge of complex models to simple models. However, the current mainstream generative large language model distillation algorithms mainly focus on optimizing the loss of the final output layer between the teacher-student models, and neglect the exploration of the model middle layer. In addition, the research on interlayer distillation often has strict requirements on the consistency of the structure of the teacher-student model, and can not deal with the distillation problem between heterogeneous models, so there are obvious limitations. To solve these problems, we propose a new knowledge distillation algorithm, which introduces the heterogeneous generative teacher-student large language model

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版
*Corresponding author.

knowledge distillation algorithm with intermediate layer distillation loss. The algorithm first extracts the middle layer information of the teacher-student model as the distillation object, and then realizes the knowledge alignment and loss calculation between heterogeneous models based on the middle layer through the specially designed middle layer mapping rules and alignment module. Finally, the ratio of each distillation loss is optimized jointly. Through experimental validation on five relevant data sets, our method shows significant advantages in improving the distillation efficiency.

**Keywords:** White Box Large Language Model , Knowledge Distillation , Intermediate Layer , Isomerism

## 1 引言

近年来，大语言模型（Large Language Model，简称LLM）蓬勃发展(Brown et al., 2020; Achiam et al., 2023; Chowdhery et al., 2023; Anil et al., 2023; Thoppilan et al., 2022)，在对话生成、翻译、问答、文本摘要等多项自然语言处理任务中均展现出了强大的性能(Zhao et al., 2023; Chang et al., 2023)。然而由于参数规模庞大，给资源受限环境下的训练和推理带来了巨大的困难(Strubell et al., 2019)。

相比之下，具有较少参数量的语言模型为资源受限的研究环境提供了一种更为可行的替代方案。但由于规模较小，其所涵盖的知识和语言能力相对有限。而知识蒸馏(Hinton et al., 2015; Adriana et al., 2015)便是一种将知识从参数规模大而复杂的深度模型转移到轻量而高效的小模型中的模型压缩技术(Deng et al., 2020; Han et al., 2015)，它通过将大模型学习到的丰富知识蒸馏到较小的模型中，使得小模型能够从大模型的优越性能中受益，平衡了小参数与高性能(Gou et al., 2021)，是大语言模型时代的一项极具发展前景的技术。

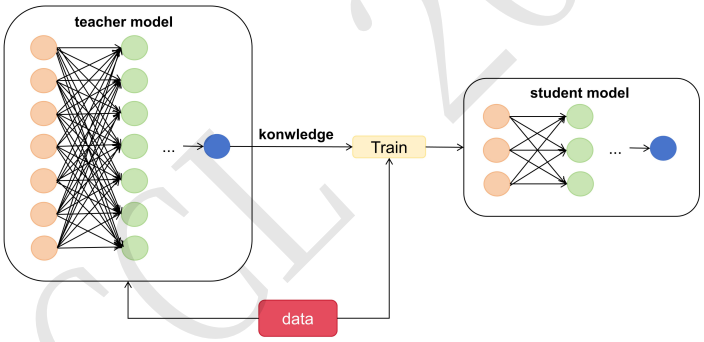



Figure 1: 知识蒸馏算法简图。知识蒸馏是将知识从一个庞大复杂的教师模型转移到一个轻量高效的学生模型中。

如图1所示，在大语言模型时代来临之前，传统的知识蒸馏研究(Adriana et al., 2015; Hinton et al., 2015)重点放在特定神经网络结构和为任务量身定做的训练目标设计上(Xu et al., 2024)。而在大语言模型时代，根据教师模型参数的可获取与否，知识蒸馏主要分为：对黑盒LLM的蒸馏(Jiang et al., 2023; Wu et al., 2023)和对白盒LLM的蒸馏(Timiryasov and Tastet, 2023; Liang et al., 2023; Gu et al., 2023)。相比于黑盒模型在蒸馏过程中只有教师预测结果可用，白盒大语言模型易于获取教师中间层结果和所有参数信息(Gou et al., 2021)。此外，白盒大语言模型根据结构的不同分为由编码器组成的语言理解模型(Devlin et al., 2018)和由解码器组成的生成式语言模型(Radford et al., 2019; Touvron et al., 2023a; Touvron et al., 2023b)。由于现有的白盒蒸馏方法主要应用于编码器结构的语言理解模型，而针对解码器结构的生成式大语言模型尚未有深度的探究，因此本文的主要研究对象是生成式大语言模型的白盒蒸馏。

现有的生成式大语言模型的白盒蒸馏方法大多仅考虑教师模型与学生模型的最后输出层蒸馏损失(Gu et al., 2023; Agarwal et al., 2023; Sanh et al., 2019; Wen et al., 2023; Kim and Rush, 2016)，忽略了对教师模型中间层信息的蒸馏。而在大语言模型的生成过程中，中间层占据了它


(c) 2024 中国中文信息学会计算语言学专业委员会

结构的一大部分，包含着许多信息：丰富的语义信息、对输入句子的注意力分配等(Clark et al., 2019; Sun et al., 2019; Wang et al., 2020b; Wang et al., 2020a)。因此，对于大语言模型中间层信息的蒸馏具有很大的研究意义。

据调研，针对Transformer架构模型的中间层蒸馏方法(Jiao et al., 2019; Sun et al., 2019; Wang et al., 2020b)往往对教师模型与学生模型的内部结构要求严格一致性：教师模型的Transformer块数需为学生的整倍数、学生模型的注意力头数及中间层隐状态维度需与教师模型完全相同等，而现实中许多蒸馏场景中教师模型与学生模型都是异构的。因此，现有的针对中间层的蒸馏算法具有明显局限性，需要一个结构无关的更普适的中间层蒸馏算法。此外，虽然针对中间层的蒸馏方法在编码器结构的语言理解模型中很常见(Sun et al., 2019; Jiao et al., 2019; Wang et al., 2020a; Hou et al., 2020; Liang et al., 2020)，但迄今为止还没有结构无关的Transformer架构的生成式大语言模型间的中间层蒸馏算法研究(Xu et al., 2024)。

为解决上述问题，我们提出了一种针对生成式大语言模型的中间层蒸馏算法，该算法放松了对教师LLM与学生LLM中间层内部结构的一致性要求，更加普适于异构生成式大语言模型之间的蒸馏场景。此外，我们将设计的中间层蒸馏算法应用于仅在输出层对教师LLM进行蒸馏的三种蒸馏算法上：标准知识蒸馏(Sanh et al., 2019; Wen et al., 2023)、序列化知识蒸馏(Kim and Rush, 2016)以及Minillm(Gu et al., 2023)，并分类地对不同的中间层输出结果进行了多项实验，最后优化蒸馏损失权重配比以最大化模型蒸馏效果。具体地，我们选用开源的GPT2系列模型(Brown et al., 2020)作为蒸馏任务中的教师与学生，其中，学生模型的参数量大小从120M（gpt2-base）到760M（gpt2-large）不等，教师模型选用参数量为1.5B的gpt2-xlarge。实验涵盖了大范围的NLP任务，使用5个指令跟随数据集：DollyEval，SelfInst(Wang et al., 2022a)，VicunaEval(Chiang et al., 2023)，S-NI(Wang et al., 2022b)以及UnNI(Honovich et al., 2022)，结合GPT4-score(Achiam et al., 2023)和Rouge-L(Lin, 2004)两种机器指标以及人工评测对蒸馏后的学生模型进行综合评估。多项实验表明，引入中间层蒸馏损失后得到的学生模型在几乎所有数据集上的表现都优于仅针对最后输出层蒸馏得到的学生模型，验证了我们提出的更普适于异构生成式LLM的中间层蒸馏算法的有效性。

## 2 相关工作

知识蒸馏（Knowledge Distillation）旨在从复杂的教师模型中提取知识训练简化的学生模型，从而使学生模型在保持相对较高性能的同时拥有更小的模型体积和更低的计算成本。

早期的知识蒸馏致力于将知识从庞大复杂的模型转移到轻量高效的网络中。Geoffrey Hinton利用教师模型的输出概率分布（Soft Target）(Hinton et al., 2015)作为学生模型的训练目标，相比于最大化正确答案的平均对数概率（Hard Target）能得到更多信息以更好地训练小网络，让小模型获得大模型的泛化能力。后来的序列级知识蒸馏（简称SeqKD）(Kim and Rush, 2016)通过最大化教师模型生成序列的可能性来微调学生模型，使学生的预测与教师的预测保持一致。这些传统的知识蒸馏思想为后续大语言模型蒸馏奠定了基础。

近年来，随着大语言模型的蓬勃发展，知识蒸馏越来越多地应用于参数量巨大的大语言模型压缩上。主要分为两类：一是针对黑盒大语言模型(Achiam et al., 2023)的蒸馏，二是针对白盒大语言模型(Radford et al., 2019; Le Scao et al., 2022; Touvron et al., 2023a; Touvron et al., 2023b; Zhang et al., 2022; Team and others, 2023)的蒸馏。其中，黑盒知识蒸馏研究在大语言模型API生成的即时响应对的小模型微调中显示出不错的结果(Jiang et al., 2023; Wu et al., 2023)。但由于黑盒蒸馏需要调用API，这些模型成本高昂，并且在某种程度上不可复制。而相比之下，白盒蒸馏为研究人员提供了一种更透明、更容易接近的方法。

现有的白盒蒸馏研究多是仅针对教师模型和学生模型的最后输出层计算蒸馏损失。标准知识蒸馏（简称KD）(Sanh et al., 2019)用标记级概率分布标记固定的序列数据集来提取教师特征知识。具体分为两步：监督学习、教师模型学习。第一步监督学习中学生模型根据标注数据进行监督学习，此过程中的损失函数为语言模型交叉熵损失。第二步教师模型学习中学生模型以教师模型的Soft Target为目标计算蒸馏损失进行训练。最后把这两个损失结果相加即为KD损失。上述标准知识蒸馏和序列化知识蒸馏方法本质上都是最小化教师和学生分布之间的近似前向KL散度：$KL(p \parallel q_\theta)$(Sanh et al., 2019; Wen et al., 2023; Timiryasov and Tastet, 2023; Liang et al., 2023)，24年提出的Minillm方法(Gu et al., 2023)揭示了这一本质，并深入研究了白盒生成式LLM的蒸馏，提出对于开放文本生成任务，让学生模型去学习教师模型的概率分布会导

致学生模型学到过多教师模型错误的判断结果。因此，为解决这个问题，Minillm提出让教师模型的输出分布去拟合学生模型的概率分布，即最小化教师和学生分布之间的近似反向KL散度：$KL(q_\theta \parallel p)$，并在整体模型的训练中加入了PPO强化学习算法(Schulman et al., 2017)，使用策略梯度优化(Sutton et al., 1999)策略选择，进一步优化了模型的蒸馏效果。

除了仅针对最后输出层的白盒蒸馏，近年来许多针对Transformer编码器架构的模型中间层蒸馏算法（简称ILD）提出，这些研究根据蒸馏的中间层对象的不同分为两类：对隐藏层的蒸馏、对自注意力模块的蒸馏。第一类中，Patient Teacher(Sun et al., 2019)不仅让学生网络从教师的Soft Target学习，还提出patient teacher-student机制，让学生模型耐心地从教师模型的多个中间层学习增量知识提取。第二类中，MiniLM(Wang et al., 2020b)提出深度自注意力精馏，通过提取教师模型最后一个Transformer层的自注意力模块让小模型通过深度模仿大模型的自注意力模块进行训练，并实验验证了对教师模型中间隐藏层蒸馏的有效性。紧随其后的MiniLMv2(Wang et al., 2020a)在此基础上进行了推广，它不再要求师生模型的自注意力头数一致，且深入研究了教师模型的中间层层次选择策略。此外，Tinybert(Jiao et al., 2019)针对BERT系列模型设计了ILD，兼具了对教师模型两种中间层对象的蒸馏，减少了模型层数但同时几乎保持了原教师模型的性能，但它对师生模型的中间层结构要求严格一致性，因此有明显的局限性。ALP-KD(Passban et al., 2021)提出了一种依赖于中间层注意力的组合蒸馏技术，它融合了教师端信息并考虑了教师模型每一层的重要性，解决了以往ILD会跳过一些教师层而仅从其子集中提取知识的缺陷。针对ILD无法轻松处理层映射搜索和师生模型架构不匹配等问题，Universal-KD(Wu et al., 2021)通过基于注意力的层投影来匹配输出空间中教师和学生的中间层（在中间层上添加伪分类器）。RAIL-KD(Haidar et al., 2021)提出了一种新的师生模型之间中间层映射方式：在每一轮训练时，随机选取教师模型的中间层与学生模型的中间层建立映射以进行知识蒸馏，其在BERT等系列教师模型的蒸馏任务中验证了通用性与有效性。最新的研究中，CR-ILD(Ko et al., 2023)发现了ILD在训练数据集上的过度拟合问题，通过仅提取教师模型最后一个Transformer层并在补充任务上进行ILD有效缓解了此种现象。针对当前的知识蒸馏方法依赖于模型的中间层特征和标签信息，因而通常需要对齐的模型架构和足够的标记数据这一问题，2023年提出的GLMD(Tan et al., 2023)方法执行两阶段单词预测蒸馏和词汇压缩，通过消除师生模型之间的维度和结构的限制以及通过没有中间层和硬标签来消除对标记数据集的需求来支持更通用的应用场景。综上，现有的白盒蒸馏方法多用于Transforme编码器架构的语言理解模型蒸馏，而对于生成式大语言模型的异构师生模型的中间层蒸馏算法尚有待探索，而这正是本文的研究主题。

## 3 模型介绍

本文算法主要包含四个模块: 中间层对象选择模块、中间层结果对齐模块、中间层损失计算模块以及各蒸馏损失融合优化模块。

整体流程框架如图2所示，其中，中间层对象选择模块阐述了我们选择的不同中间层结果种类及其含义；中间层结果对齐模块是我们算法的核心，它实现了异构师生模型中间层结果的对齐。主要包含两个部分：一是中间层映射规则设计，这部分详细介绍了我们为不同层数关系的教师与学生模型设计的层对应规则；二是维度转换模块，这部分说明了当面对中间层输出结果维度不一致时，我们是如何进行维度转换的；中间层损失计算模块讲述了我们对最终损失函数的设计；最后的各蒸馏损失融合优化模块介绍了我们是如何调节各部分蒸馏的损失权重配比，以达到优化蒸馏效果这一目标的。

### 3.1 中间层蒸馏对象选择

考虑到基于Transformer架构的开源大语言模型的中间层输出结果主要包含了两类：注意力模块输出、隐藏层输出。这里隐藏层输出包括输入模型主体部分前的词嵌入层、模型每个Transformer层的中间隐藏层输出以及模型最后一层输出。实验中，我们仅使用了每个Transformer Layer的中间隐藏层输出作为蒸馏对象。考虑到这些中间层输出结果代表的含义不同，蕴含的信息也不同，在中间层信息上的重要程度占比未知，蒸馏效果也有待考量(Clark et al., 2019)。因此，我们针对以上两种不同类型的中间层结果，分别进行了蒸馏实验。使用两种不同的中间层输出作为蒸馏对象后得到的实验结果对比见表3,表4,表5。


(c) 2024 中国中文信息学会计算语言学专业委员会

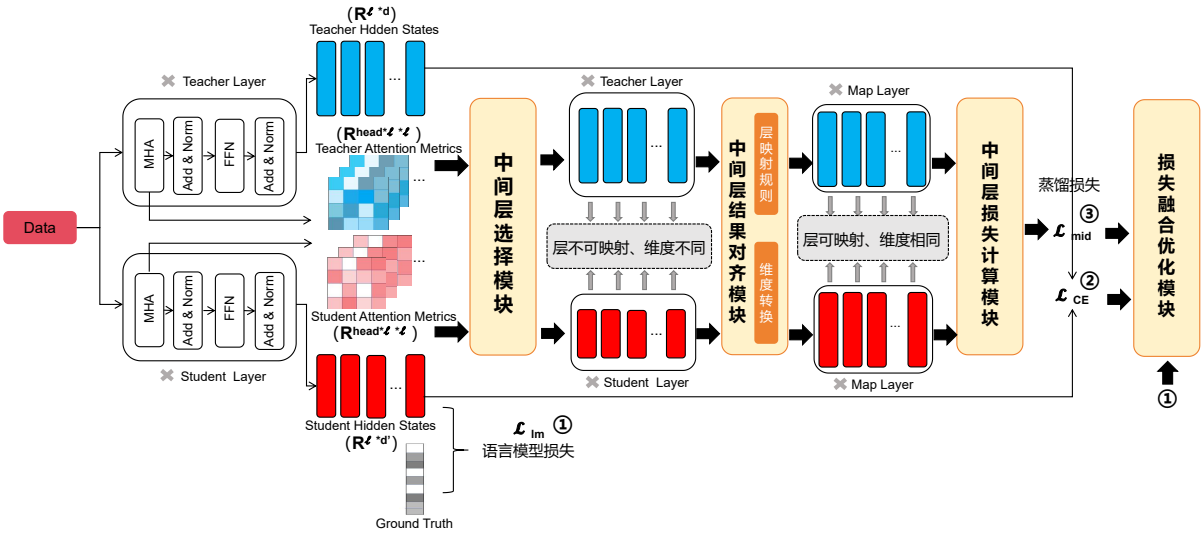



Figure 2: 引入中间层损失的异构师生模型蒸馏算法框架图。包括中间层对象选择、中间层结果对齐、中间层损失计算、各蒸馏损失融合优化四个模块。

### 3.2 中间层结果对齐模块

与Tinybert(Jiao et al., 2019)不同的是，本文研究的教师模型与学生模型的中间层输出在结构上不同，主要体现在组成教师模型和学生模型的Transformer块的隐藏层状态维度不同、多头注意力模块中的注意力头数不同、中间层Transformer块的个数不成倍数关系等结构细节上，如图3。以上原因直接导致了在中间层蒸馏的实现过程中不能直接利用教师模型和学生模型的中间层输出结果进行蒸馏损失计算，而要首先进行中间层结果的对齐。为此，我们设计了更普适的异构师生模型中间层映射规则和维度转换模块。

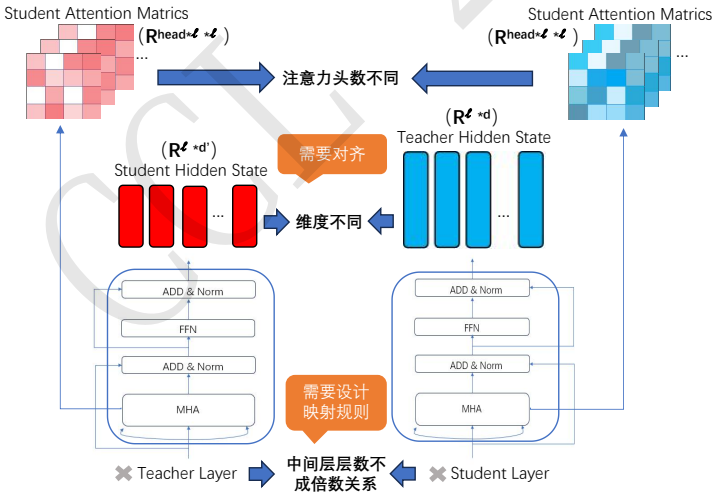



Figure 3: 师生模型中间层异构性示意图。展示了本文研究的师生模型在中间结构上的异构细节

#### 3.2.1 中间层映射规则

具体的，层映射规则设计为:取教师模型与学生模型的中间层层数的最大公因数为后续中间层对齐时的Transformer组合块数，若最大公因数不为1，则最大公因数的值即为教师模型和学生模型划分出的中间层组合块数，为后续的中间层蒸馏一一对应。当遇到最大公因数为1的情况时，就依次将教师模型、学生模型的中间层数循环减1，直到最大公因数不为1时停止循环，剩余的中间层按照上述规则进行组合对齐，减掉的部分中间层也同样相应对齐。举个例子，当教

师模型中间层为48层，学生模型为20层时，属于情况一：取师生模型层数的最大公因数4，将教师模型和学生模型的中间层都分别划分为四个大块，这样教师模型每块有12层，学生模型每块有5层，使用以块为单位的中间输出结果作为蒸馏对象进行后续的中间层蒸馏损失计算。而当教师模型为49层，学生模型为19层时，属于情况二，则循环依次将教师模型、学生模型层数减一，直到师生模型的层数最大公因数不为1停止循环，操作后剩余的层之间和减去的对应层之间均参考情况一的处理方法，这样设计是为了后续能层次对应均匀地进行中间层蒸馏。这部分的可解释性实验见5.2.2，具体的处理如下伪代码:

**Algorithm 1** 师生模型中间层映射规则

**Input:** 教师模型中间层注意力矩阵$t_atts$, 学生模型中间层注意力矩阵$s_atts$
**Output:** 教师模型最终映射矩阵$map_t$, 学生模型最终映射矩阵$map_s$

$max\_gcd \leftarrow$ 师生模型中间层层数的最大公因数
**while** $max\_gcd == 1$ **do**
    **if** 第偶数次（初始为0） **then**
        教师中间层层数减一
    **else**
        学生中间层层数减一
    **end if**
    $max\_gcd \leftarrow$ 当下师生模型中间层层数的最大公因数
**end while**
获取教师模型中间层参与映射的总块数$t_block \leftarrow \lfloor t_layer/max_gcd \rfloor$
获取学生模型中间层参与映射的总块数$s_block \leftarrow \lfloor s_layer/max_gcd \rfloor$
计算教师模型最终的中间层映射矩阵$map_t \leftarrow [t_atts[i \times t_block + t_block - 1]$
计算学生模型最终的中间层映射矩阵$map_s \leftarrow [s_atts[i \times s_block + s_block - 1]$
**if** 属于情况2 **then**
    被减去的师生模型中间层部分按照同样的规则得到对应映射，并加入最终映射矩阵中
**else**
    保持原样
**end if**
最终输出$map_t, map_s$

### 3.2.2 维度转换模块

中间层映射规则确定后，就需要统一教师模型与学生模型的中间层输出维度了，让他们可以进行蒸馏损失值的计算。具体的，我们在模型层面设计了一个线性映射层，负责将教师模型的中间层结果通过这一线性映射网络转换为与学生模型的中间层输出结果维度一致的张量，然后进行中间层损失计算。这个线性映射网络与学生模型一起参与训练优化，尽可能地让维度转换后的教师模型中间层结果保留其语义信息，以传递给学生模型。最终的中间层蒸馏目标如下：

$$\mathcal{L}_{\text{mid}} = \sum_{i=1}^{N} \mathcal{L}_{\text{F}}(\Phi(f_i^T(x)), f_i^S(x)) \tag{1}$$

其中，N指进行蒸馏的中间层对应映射的层数，i的取值取决于3.2.1中讲述的中间层映射规则，$f_i^T(x)$和$f_i^S(x)$分别是教师模型和学生模型第i个Transformer块的中间层输出结果，mid与$f$的种类取决于3.1中的中间层蒸馏对象选择，转换函数$\Phi_T$应用于教师模型的中间层,进行维度转换，以确保教师模型与学生模型的中间层输出具有相同的形状，便于直接比较，而损失函数类型$\mathcal{L}_{\text{F}}$将在3.3中选择与确定。

## 3.3 中间层损失计算模块

在中间层损失函数的选择过程中，为了与标准知识蒸馏方法中的仅输出层的蒸馏损失$\mathcal{L}_{\text{CE}}$与语言模型损失$\mathcal{L}_{\text{lm}}$在数值量级上保持对齐。我们尝试了多种不同的中间层损失函数来对教师模型和学生模型进行中间层输出结果的损失计算。具体地，我们尝试了余弦相似性损


(c) 2024 中国中文信息学会计算语言学专业委员会

失函数（最大化师生模型中间层输出张量的相似度）、平均绝对误差MAE（最小化师生模型中间结果的差）、均方误差MSE（最小化师生模型中间层输出结果的平方差），结果对比见表7，最终根据实际实验效果选择了MSE作为损失函数，因此最终的中间层损失函数为：

$$\mathcal{L}_{\text{mid}} = \frac{1}{N}\sum_{i=1}^{N}(\Phi(f_i^T(x)) - f_i^S(x))^2 \tag{2}$$

### 3.4 各蒸馏损失融合优化模块

在获取了中间层蒸馏损失后，我们设计了各部分损失的融合优化模块，包括：语言模型损失$\mathcal{L}_{\text{lm}}$、最后输出层的蒸馏损失$\mathcal{L}_{\text{CE}}$、我们引入的中间层蒸馏损失$\mathcal{L}_{\text{mid}}$，各部分的损失计算公式如下：

语言模型损失:

$$\mathcal{L}_{\text{lm}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{i,j}\log(p_{i,j}) \tag{3}$$

其中，N是指样本数量，C是类别数量（词汇表大小），$y_{i,j}$是样本i的真实标签，$p_{i,j}$是模型对于样本i中第j类别的预测概率。

最后输出层的蒸馏损失:

$$\mathcal{L}_{\text{CE}} = \sum_i t_i \log(s_i) \tag{4}$$

其中,$t_i$是指教师模型的预测分布，$s_i$是指学生模型的预测分布，两者计算交叉熵损失。

引入的中间层蒸馏损失:

$$\mathcal{L}_{\text{mid}} = \begin{cases} \frac{1}{N}\sum_{i=1}^{N} MSE(\Phi_{attn}(A_i^T), A_i^S), & mid = attention \\ \frac{1}{N}\sum_{i=1}^{N} MSE(\Phi_{layer}(o_i^T(x)), o_i^S(x)), & mid = hidden\ layer \end{cases} \tag{5}$$

其中，$MSE$是指MSE损失函数，$\Phi_{attn}$（$\Phi_{layer}$）是指针对注意力模块（隐藏层）设计的线性映射，$o_i(x)$是指中间隐藏层输出，$A_i$是指中间注意力模块输出。

我们使用网格搜索方法根据模型蒸馏效果自动优化配比三种不同损失在总损失上所占比重，最终得到融合了三种不同损失的优化目标，公式如下：

$$\mathcal{L}_{\text{sum}} = \alpha\mathcal{L}_{\text{lm}} + \beta\mathcal{L}_{\text{CE}} + \sum_{i=0}^{N}\lambda_i\mathcal{L}_{\text{layer}} \tag{6}$$

其中，N指进行蒸馏的中间层对应映射的层数，$\mathcal{L}_{\text{layer}}$是第i层的中间层蒸馏损失，$\lambda_i$是第i层的蒸馏损失占整个中间层蒸馏中的重要性参数，$\alpha$和$\beta$都是损失重要性系数。

## 4 实验设置

### 4.1 实验训练设置

实验选用的蒸馏数据集来自$databricks-dolly-15k$[1]，由15K对人类书写的指令-响应组成。我们使用其中随机分割的14K个样本作为训练集$D$，剩下的500个作为验证集和测试集。对于模型预训练数据集$\mathcal{D}_{\text{PT}}$，我们使用OpenWebText(Gokaslan et al., 2019)。实验过程中，我们首先在$D$上微调一个生成式大语言模型作为教师$p$，然后通过评估蒸馏模型在五个指令跟随数据集上的表现比较不同知识蒸馏方法的蒸馏效果。实验流程与训练细节如下：

- **实验流程**：我们首先按照4.1节中所述选取并划分数据集，接着参考之前的研究(Gu et al., 2023)在预处理好的指令-回复数据集$D$上微调大语言模型GPT2-xlarge作为后续蒸馏的教师模型；然后分别对三种学生模型GPT2-base、GPT2-medium、GPT2-large在数据集$D$上使用不同的蒸馏方案蒸馏：仅在$D$上使用标签数据监督微调（SFT w/o KD）、标准蒸馏（KD）、序列化蒸馏（SeqKD）、MiniLLM方法、引入中间层蒸馏损失的蒸馏方法；比较这些不同的知识蒸馏方法蒸馏训练后得到的学生模型在预先划分出来的测试集上生成的回复内容的Rouge-L与GPT4得分，并最终评估蒸馏效果的好坏。

[1]https://github.com/databrickslabs/dolly/tree/master


(c) 2024 中国中文信息学会计算语言学专业委员会

- **训练与参数设置**：对于基线方法，学习率取值区间设为[5e-4,1e-4,5e-5]，总批大小取值区间设为[32,64]，蒸馏训练迭代20轮。对于标准知识蒸馏（KD），以0.5的混合率将蒸馏损失与语言模型损失混合得到最终损失。每个基线方法的检查点由验证集的Rouge-L得分选择，选择分数最高的为最优的蒸馏模型。对于MiniLLM方法，首先使用数据集D对学生模型进行监督微调后选择验证损失最小的检查点作为后续MiniLLM训练的初始模型，使用Adamw优化器，学习率设为5e-6，总批大小设为64，采样时温度设为1，模型允许的最大文本长度为512，同样的，使用验证集上的Rouge-L分数选择最后的检查点。对于引入中间层蒸馏损失的方法，本文提出的方法分别加在了上述基线方法中进行中间层蒸馏实验，因此超参数设置与上述设置一致，除了最后得到的中间层蒸馏损失在最终的损失函数占比上的权重配比使用网格搜索算法进行了优化，各损失的权重搜索范围设为[0,1]，选择0.1的间隔进行搜索，使用交叉验证的平均损失值作为评价指标，遍历所有可能的权重组合以寻找最佳的权重配比方案。以上实验基于8卡NVIDIA RTX A5000，并使用deepspeed进行模型分布式训练。

### 4.2 对比方法

为了评估在生成式大语言模型中引入中间层蒸馏的有效性，我们与以下基线方法作对比：

- **SFT w/o KD**：直接在数据集$D$上进行监督微调，不使用知识蒸馏。
- **KD**(Sanh et al., 2019)：词级别的知识蒸馏。使用教师分布作为每个token级别的监督，在数据集$D$上对学生模型进行微调，最小化师生模型的正向KL散度。
- **SeqKD**(Kim and Rush, 2016)：使用教师模型生成的数据对学生模型进行微调。
- **MiniLLM**(Gu et al., 2023)：类似于KD，不同点在于最小化师生模型的反向KL散度，且加入了模型优化策略。

### 4.3 评测指标

我们使用以下两个指标评估模型生成的响应：

- **Rouge-L**(Lin, 2004)：用来评估模型生成内容的精度，(Wang et al., 2022b)表明Rouge-L适用于大规模指令跟踪评估。具体地，Rouge-L指标结果是对各个蒸馏方法在每个数据集的五个不同随机种子下生成的回答内容计算Rouge-L值最后取平均得到的。
- **GPT4**(Achiam et al., 2023)：设计prompt将蒸馏模型生成的结果与ground truth标签数据进行比较打分，分值为1-10分，关于使用GPT4评分时的prompt设计详见附录A。具体地，对于每个蒸馏模型，我们使用五个不同的随机种子生成回复，最后的GPT4打分结果是将各个蒸馏方法在各个数据集的不同随机种子设置下生成的结果均匀采样100条数据后调用GPT4接口对回复内容进行打分后获得的分数取平均值得到的。

## 5 实验结果与分析

### 5.1 对比实验

我们复现了KD、SeqKD、MiniLLM这三种仅在模型最后输出层蒸馏的算法作对照，分别在这些方法上引入了我们提出的中间层蒸馏算法进行实验，并在DollyEval、SelfInst、Vicuna、S-NI、UnNI五个指令跟随数据集上进行了蒸馏模型的效果评测,包括机器评测和人工评测。


(c) 2024 中国中文信息学会计算语言学专业委员会

| Model | Params | Method | DollyEval | | SelfInst | | VicunaEval | | S-NI | | UnNI | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | R-L | GPT4 | R-L | GPT4 | R-L | GPT4 | R-L | GPT4 | R-L | GPT4 |
| GPT2 | 1.5B | Teacher | 27.1 | 60.9 | 14.7 | 55.9 | 16.4 | 64.0 | 27.3 | 40.5 | 31.5 | 69.0 |
| | 120M | SFT w/o KD | 23.0 | 41.2 | 10.4 | 36.4 | 14.2 | 35.5 | 16.5 | 25.0 | 18.4 | 34.3 |
| | | KD | 22.4 | 20.6 | 9.4 | 28.3 | 13.7 | 33.2 | 15.0 | 25.8 | 15.5 | 39.4 |
| | | **OurMethod** | **23.6** | **20.9** | **9.8** | **33.0** | **15.0** | **36.1** | **15.8** | **30.5** | **18.0** | 32.8 |
| | | SeqKD | 17.8 | 13.4 | 7.3 | 17.1 | 11.9 | 22.9 | 9.5 | 16.5 | 11.0 | 27.6 |
| | | **OurMethod** | **18.4** | **14.7** | **8.5** | **23.9** | **12.5** | **23.7** | **12.6** | **20.3** | **13.6** | 26.7 |
| | | Minillm | 19.3 | 17.4 | 9.7 | 31.3 | 11.9 | 36.4 | 17.0 | 25.9 | 18.7 | 31.0 |
| | | **OurMethod** | 19.0 | **22.2** | **9.8** | **35.9** | **12.2** | **38.8** | 16.9 | **29.8** | **19.2** | **37.6** |
| | 340M | SFT w/o KD | 25.2 | 61.7* | 12.2 | 42.3 | 16.1 | 49.4 | 20.9 | 49.7* | 24.5 | 33.9 |
| | | KD | 22.2 | 43.1 | 11.0 | 42.4 | 14.2 | 39.6 | 18.5 | 32.2 | 21.0 | 36.2 |
| | | **OurMethod** | **25.5** | **53.4** | **12.3** | **46.3** | **15.7** | **43.5** | **21.1** | **37.2** | **25.9** | **40.4** |
| | | SeqKD | 25.0 | 55.6 | 12.5 | 49.4 | 16.0 | 46.4 | 23.1 | 35.6 | 26.9 | 43.8 |
| | | **OurMethod** | **25.4** | **58.2** | **12.7** | 45.3 | **16.8*** | **46.5** | 23.1 | **43.8*** | **27.4** | **47.5** |
| | | Minillm | 21.1 | 46.8 | 11.9 | 40.6 | 12.6 | 44.8 | 20.0 | 22.8 | 23.8 | 47.5 |
| | | **OurMethod** | **22.5** | **50.1** | **12.2** | **43.4** | **14.1** | **45.5** | **22.8** | **35.6** | **26.3** | **48.3** |
| | 760M | SFT w/o KD | 25.7 | 61.6* | 12.1 | 39.6 | 16.5* | 45.5 | 21.2 | 44.5* | 24.9 | 55.6 |
| | | KD | 24.9 | 57.0 | 12.0 | 40.7 | 15.9 | 48.4 | 20.8 | 41.8* | 23.5 | 31.5 |
| | | **OurMethod** | **25.6** | **59.7** | **12.8** | **45.6** | **16.2** | 39.3 | **23.3** | **44.5*** | **26.0** | **35.2** |
| | | SeqKD | 25.1 | 58.0 | 12.6 | 36.9 | 16.3 | 42.2 | 21.3 | 42.0* | 23.9 | 24.2 |
| | | **OurMethod** | **26.1** | **60.1** | 12.5 | **44.0** | **16.4*** | **42.7** | **22.5** | 41.6* | **26.2** | **30.0** |
| | | Minillm | 21.8 | 58.9 | 13.2 | 57.5* | 13.8 | 52.2 | 23.3 | 55.1* | 25.5 | 33.3 |
| | | **OurMethod** | **23.1** | **60.4** | **14.6** | 54.0 | **15.9** | **53.5** | **25.6** | **58.3*** | **26.3** | **39.8** |

Table 1: 机器评测结果。GPT4和R-L分别代表GPT-4和Rouge-L在5个随机种子中的平均反馈分数。加入我们的方法后得分有提升的加粗表示，学生模型的表现优于教师的标记*。

| Method | DollyEval | SelfInst | VicunaEval | S-NI | UnNI |
|---|---|---|---|---|---|
| KD | 4.33 | 5.53 | 4.32 | 4.73 | 3.16 |
| SeqKD | 3.98 | 4.74 | 4.21 | 5.2 | 2.63 |
| MiniLLM | 5.08 | 6.94 | 5.4 | 6.67 | 3.54 |
| **OurMethod** | **5.13** | **7.12** | **6.71** | **6.75** | **4.42** |

Table 2: 人工评测结果。分值范围为1-10分，结果是随机选取指定方法在指定数据集下的生成内容打分后取平均得到的。

### 5.1.1 机器评测结果

使用Rouge-L和GPT4-score这两个机器评价指标得到的结果如表1。根据结果分析可得：首先，相比于监督微调（SFT w/o KD），引入本文提出的中间层蒸馏损失后的三种基线方法均在大多数情况下获得了更高的RougeL和GPT4分数，且在相同规模的师生蒸馏场景中基本获得了所有蒸馏方案中最高的分数，成功实现了对教师模型的知识蒸馏效果的优化，验证了本文提出算法的有效性。其次，通过比较引入我们提出的中间层蒸馏算法后的三种基线方法（OurMethod）与未引入的基线方法（KD、SeqKD、MiniLLM）在五个数据集上的的各指标得分，引入中间层蒸馏损失后的得分普遍获得了提升，进一步验证了本文方法有效提取了输出层之外额外的教师模型信息（中间层），进而实现了蒸馏效果的提升，也说明了本方法对仅输出层蒸馏方法的普适性、可叠加性。

### 5.1.2 人工评测结果

针对每种蒸馏方法，我们随机抽取了部分生成样本进行人工打分(1-10)，最后取样本平均


(c) 2024 中国中文信息学会计算语言学专业委员会

分为每种蒸馏算法得分。打分过程中，以ground truth为标准答案，并综合考虑生成内容的流畅性、完整性、正确性以及内容多样性等因素。人工评测结果如表2，可以看出，相对于仅针对输出层蒸馏的算法，引入了中间层损失的蒸馏方法获得了更高的分数，说明使用我们的算法蒸馏得到的学生模型在语言流畅性、语义完整性和灵活性等角度性能得到了提升。

## 5.2 分析实验

### 5.2.1 不同中间层对象的蒸馏实验结果

在中间层蒸馏细节上，我们分别利用了两种不同的中间层蒸馏对象进行实验：中间层注意力输出、隐藏层输出。综合表3,4,5的各评测指标得分可得，无论是何种中间层蒸馏对象，引入了中间层蒸馏损失之后，原基线蒸馏方法的指标得分均得到了一定的提升，进一步证明了本文提出的中间层蒸馏算法能有效提升异构生成式大语言模型的蒸馏效果。此外，相较于选择隐藏层输出作为蒸馏对象，引入中间层注意力模块的蒸馏损失在多数情况下得到的学生模型得分更高，蒸馏效果更好，也一定程度上说明了注意力模块可能蕴含更多的中间语义有效信息。

| Model | Params | Method | DollyEval | | SelfInst | | VicunaEval | | S-NI | | UnNI | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | R-L | GPT4 | R-L | GPT4 | R-L | GPT4 | R-L | GPT4 | R-L | GPT4 |
| GPT2 | 1.5B | Teacher | 27.1 | 60.9 | 14.7 | 55.9 | 16.4 | 64.0 | 27.3 | 40.5 | 31.5 | 69.0 |
| | 120M | KD | 22.4 | 20.6 | 9.4 | 28.3 | 13.7 | 33.2 | 15.0 | 25.8 | 15.5 | 39.4 |
| | | **KD+attn** | **23.6** | **20.9** | **9.8** | **28.4** | **15.0** | **34.9** | **15.8** | **30.5** | **18.0** | 31.6 |
| | | **KD+layer** | **23.1** | **20.7** | **9.6** | **33.0** | **14.5** | **36.1** | **15.2** | **27.0** | **17.2** | 32.8 |
| | 340M | KD | 22.2 | 43.1 | 11.0 | 42.4 | 14.2 | 39.6 | 18.5 | 32.2 | 21.0 | 36.2 |
| | | **KD+attn** | **25.5** | **53.4** | **12.3** | **46.3** | **15.7** | **43.5** | **21.1** | **32.7** | **25.9** | **40.4** |
| | | **KD+layer** | **24.4** | **50.9** | **11.8** | **43.5** | **14.3** | **41.4** | **20.2** | **37.2** | **25.1** | **39.8** |
| | 760M | KD | 24.9 | 57.0 | 12.0 | 40.7 | 15.9 | 48.4 | 20.8 | 41.8 | 23.5 | 31.5 |
| | | **KD+attn** | **25.6** | **58.2** | **12.8** | **44.5** | **16.0** | 39.3 | **23.3** | **44.5** | **26.0** | **35.2** |
| | | **KD+layer** | **25.2** | **59.7** | **12.5** | **45.6** | **16.2** | 39.2 | **21.2** | 38.2 | **24.4** | **33.9** |

Table 3: 不同中间层蒸馏对象应用于KD实验结果表。其中加粗表示的部分是指我们的方法超过原KD方法。

| Model | Params | Method | DollyEval | | SelfInst | | VicunaEval | | S-NI | | UnNI | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | R-L | GPT4 | R-L | GPT4 | R-L | GPT4 | R-L | GPT4 | R-L | GPT4 |
| GPT2 | 1.5B | Teacher | 27.1 | 60.9 | 14.7 | 55.9 | 16.4 | 64.0 | 27.3 | 40.5 | 31.5 | 69.0 |
| | 120M | SeqKD | 17.8 | 13.4 | 7.3 | 17.1 | 11.9 | 22.9 | 9.5 | 16.5 | 11.0 | 27.6 |
| | | **SeqKD+attn** | **18.4** | **14.7** | **8.5** | **22.4** | **12.5** | **23.1** | **12.6** | **20.3** | **13.6** | 26.7 |
| | | **SeqKD+layer** | **18.0** | **13.6** | **7.7** | **23.9** | **12.2** | **23.7** | **9.7** | **17.3** | **11.2** | 24.8 |
| | 340M | SeqKD | 25.0 | 55.6 | 12.5 | 49.4 | 16.0 | 46.4 | 23.1 | 35.6 | 26.9 | 43.8 |
| | | **SeqKD+attn** | **25.3** | **58.2** | **12.7** | 44.9 | **16.2** | **46.5** | 22.9 | **43.8** | **27.4** | **44.3** |
| | | **SeqKD+layer** | **25.4** | **57.7** | **12.7** | 45.3 | **16.8** | **46.5** | **23.2** | **37.3** | **27.1** | **47.5** |
| | 760M | SeqKD | 25.1 | 58.0 | 12.6 | 36.9 | 16.3 | 42.2 | 21.3 | 42.0 | 23.9 | 24.2 |
| | | **SeqKD+attn** | **26.1** | **59.7** | **12.8** | **44.0** | **16.4** | **42.7** | **22.5** | 41.0 | **26.2** | **30.0** |
| | | **SeqKD+layer** | **25.8** | **60.1** | 12.2 | **40.3** | 16.3 | **42.5** | **21.6** | 41.6 | **24.5** | **25.8** |

Table 4: 不同中间层蒸馏对象应用于SeqKD实验结果表。

### 5.2.2 不同中间层映射规则实验结果

在中间层映射规则的设计上本文主要从两大角度考虑，最后选择最大公因数异构对齐策略：一是参考了之前工作中已有实验验证有效性的倍数映射策略(Jiao et al., 2019)，二是在实际实验中，我们设计了纯倍数映射(Jiao et al., 2019)、随机映射(Haidar et al., 2021)、指定层映射等多种不同的映射策略，具体地，它们的含义是：

| Model | Params | Method | DollyEval | | SelfInst | | VicunaEval | | S-NI | | UnNI | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | R-L | GPT4 | R-L | GPT4 | R-L | GPT4 | R-L | GPT4 | R-L | GPT4 |
| GPT2 | 1.5B | Teacher | 27.1 | 60.9 | 14.7 | 55.9 | 16.4 | 64.0 | 27.3 | 40.5 | 31.5 | 69.0 |
| | 120M | Minillm | 19.3 | 17.4 | 9.7 | 31.3 | 11.9 | 36.4 | 17.0 | 25.9 | 18.7 | 31.0 |
| | | **Minillm+attn** | **19.5** | **17.9** | **9.8** | **35.9** | **12.2** | **38.8** | 16.9 | **26.6** | **19.2** | **37.6** |
| | | **Minillm+layer** | 19.0 | **22.2** | **9.8** | **35.2** | **12.0** | **38.7** | 16.9 | **29.8** | **18.8** | **36.3** |
| | 340M | Minillm | 21.1 | 46.8 | 11.9 | 40.6 | 12.6 | 44.8 | 20.0 | 22.8 | 23.8 | 47.5 |
| | | **Minillm+attn** | **22.5** | **50.1** | **12.2** | **43.4** | **14.1** | **45.5** | **22.8** | **24.4** | **26.3** | **48.4** |
| | | **Minillm+layer** | **22.4** | **49.6** | **12.4** | **42.8** | **13.1** | **45.2** | **22.5** | **35.6** | **26.1** | **47.9** |
| | 760M | Minillm | 21.8 | 58.9 | 13.2 | 57.5 | 13.8 | 52.2 | 23.3 | 55.1 | 25.5 | 33.3 |
| | | **Minillm+attn** | **23.1** | **59.3** | **14.6** | 52.4 | **15.9** | **53.5** | **25.6** | **58.3** | **26.3** | **33.4** |
| | | **Minillm+layer** | **22.5** | **60.4** | **13.6** | **57.8** | **15.9** | **53.3** | **25.0** | **56.7** | **25.8** | **39.8** |

Table 5: 不同中间层蒸馏对象应用于MiniLLM实验结果表。

- **纯倍数映射**：仅适用于教师模型与学生模型的中间层层数成整数倍的情况。是本文设计的最大公因数异构对齐策略中的一个特殊情况。因此实验中没有单独测试。
- **随机映射**：在每一轮训练时，随机选取教师模型的中间层与学生模型的中间层建立映射以进行知识蒸馏。
- **指定层映射**：手动规定教师模型的哪几个中间层对应学生模型的哪几个中间层，训练前指定，不同于随机映射，这里每轮训练所使用的层映射规则不变。
- **最大公因数异构对齐策略**：详见3.2.1。

我们选取教师模型为GPT2-xlarge，学生模型为GPT2-base，蒸馏方法为标准蒸馏（KD）上加入本文提出的中间层蒸馏算法，选取数据集DollyEval，分别应用上述不同的中间层映射规则开展蒸馏实验，得到的最终实验结果对比如表6所示，因此最终选择了效果最好的最大公因数异构对齐策略。

Table 6: 中间层映射规则蒸馏效果对比。加粗表示的为蒸馏效果最好的方案。

| 中间层映射规则 | Rouge-L(↑) | GPT4-score(↑) |
|---|---|---|
| 随机映射 | 22.9847 | 18.66 |
| 指定层映射 | 16.9077 | 15.14 |
| 最大公因数异构对齐策略 | **23.6245** | **20.86** |

### 5.2.3 不同的中间层损失函数实验结果

为确定3.3节的中间层损失计算模块中使用何种损失函数计算中间层蒸馏损失，我们分别应用余弦相似性Cosine Similarity、平均绝对误差MAE以及均方误差MSE三种损失函数，以GPT2-base作为学生模型，GPT2-xlarge作为教师模型，得到的Rouge-L实验效果对比如表7，因此，最终选择了效果最好的MSE作为中间层蒸馏损失函数。

| Loss Function | Method | | |
|---|---|---|---|
| | KD+OurMethod | SeqKD+OurMethod | MiniLLM+OurMethod |
| Cosine | 22.1 | 17.2 | 17.8 |
| MAE | 17.4 | 12.4 | 15.6 |
| MSE | **23.6** | **18.4** | **19** |

Table 7: 不同中间层蒸馏损失函数结果表。加粗的是Rouge-L指标最高的


(c) 2024 中国中文信息学会计算语言学专业委员会

## 6 总结

总的来说，本文提出了一种针对生成式大语言模型的在仅输出层蒸馏之外引入中间层损失的蒸馏方法，实现上主要包括四部分:中间层蒸馏对象选择、中间层结果对齐、中间层损失计算以及各损失函数优化融合,并将此中间层蒸馏方法融入多种现有的仅输出层蒸馏的白盒生成式大语言模型蒸馏方法中。我们首次将中间层蒸馏思想应用于同为Transformer解码器架构但中间层内部结构不同的师生模型中，设计了一种更加普适于异构师生模型,蒸馏效果更好的白盒生成式大语言模型蒸馏方法。实验中，我们提出的蒸馏方法在五个数据集三种不同学生模型规模的多个实验中均获得了超过原基线蒸馏方法的效果，验证了我们设计的引入中间层蒸馏方法的有效性，为未来针对生成式大语言模型的异构师生模型间的中间层蒸馏提供了思路与技术借鉴。

## 7 致谢

该研究得到了国家自然科学基金（62106105）、CCF-百度松果基金(CCF-Baidu202307)、南京航空航天大学科研启动基金（YQR21022）和南京航空航天大学高性能计算平台的支持。

## 参考文献

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Romero Adriana, Ballas Nicolas, K Samira Ebrahimi, Chassang Antoine, Gatta Carlo, and Bengio Yoshua. 2015. Fitnets: Hints for thin deep nets. *Proc. ICLR*, 2(3):1.

Rishabh Agarwal, Nino Vieillard, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. 2023. Gkd: Generalized knowledge distillation for auto-regressive sequence models. *arXiv preprint arXiv:2306.13649*.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*.

Lei Deng, Guoqi Li, Song Han, Luping Shi, and Yuan Xie. 2020. Model compression and hardware acceleration for neural networks: A comprehensive survey. *Proceedings of the IEEE*, 108(4):485–532.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Aaron Gokaslan, Vanya Cohen, Ellie Pavlick, and Stefanie Tellex. 2019. Openwebtext corpus.

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819.


(c) 2024 中国中文信息学会计算语言学专业委员会

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2023. Minillm: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*.

Md Akmal Haidar, Nithin Anchuri, Mehdi Rezagholizadeh, Abbas Ghaddar, Philippe Langlais, and Pascal Poupart. 2021. Rail-kd: Random intermediate layer mapping for knowledge distillation. *arXiv preprint arXiv:2109.10164*.

Song Han, Huizi Mao, and William J Dally. 2015. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2022. Unnatural instructions: Tuning language models with (almost) no human labor. *arXiv preprint arXiv:2212.09689*.

Lu Hou, Zhiqi Huang, Lifeng Shang, Xin Jiang, Xiao Chen, and Qun Liu. 2020. Dynabert: Dynamic bert with adaptive width and depth. *Advances in Neural Information Processing Systems*, 33:9782–9793.

Yuxin Jiang, Chunkit Chan, Mingyang Chen, and Wei Wang. 2023. Lion: Adversarial distillation of closed-source large language model. *arXiv preprint arXiv:2305.12870*.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2019. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351*.

Yoon Kim and Alexander M Rush. 2016. Sequence-level knowledge distillation. *arXiv preprint arXiv:1606.07947*.

Jongwoo Ko, Seungjoon Park, Minchan Jeong, Sukjin Hong, Euijai Ahn, Du-Seong Chang, and Se-Young Yun. 2023. Revisiting intermediate layer distillation for compressing language models: An overfitting perspective. *arXiv preprint arXiv:2302.01530*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model.

Kevin J Liang, Weituo Hao, Dinghan Shen, Yufan Zhou, Weizhu Chen, Changyou Chen, and Lawrence Carin. 2020. Mixkd: Towards efficient distillation of large-scale language models. *arXiv preprint arXiv:2011.00593*.

Chen Liang, Simiao Zuo, Qingru Zhang, Pengcheng He, Weizhu Chen, and Tuo Zhao. 2023. Less is more: Task-aware layer-wise distillation for language model compression. In *International Conference on Machine Learning*, pages 20852–20867. PMLR.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Peyman Passban, Yimeng Wu, Mehdi Rezagholizadeh, and Qun Liu. 2021. Alp-kd: Attention-based layer projection for knowledge distillation. In *Proceedings of the AAAI Conference on artificial intelligence*, volume 35, pages 13657–13665.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in nlp. *arXiv preprint arXiv:1906.02243*.

Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. 2019. Patient knowledge distillation for bert model compression. *arXiv preprint arXiv:1908.09355*.


(c) 2024 中国中文信息学会计算语言学专业委员会

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. 1999. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12.

Shicheng Tan, Weng Lam Tam, Yuanchun Wang, Wenwen Gong, Shu Zhao, Peng Zhang, and Jie Tang. 2023. Are intermediate layers and labels really necessary? a general language model distillation method. *arXiv preprint arXiv:2306.06625*.

MosaicML NLP Team et al. 2023. Introducing mpt-7b: A new standard for open-source, commercially usable llms, 2023. *URL www. mosaicml. com/blog/mpt-7b. Accessed*, pages 05–05.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Inar Timiryasov and Jean-Loup Tastet. 2023. Baby llama: knowledge distillation from an ensemble of teachers trained on a small dataset with no performance penalty. *arXiv preprint arXiv:2308.02019*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. 2020a. Minilmv2: Multi-head self-attention relation distillation for compressing pretrained transformers. *arXiv preprint arXiv:2012.15828*.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020b. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022a. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. 2022b. Benchmarking generalization via in-context instructions on 1,600+ language tasks. *arXiv preprint arXiv:2204.07705*, 2.

Yuqiao Wen, Zichao Li, Wenyu Du, and Lili Mou. 2023. f-divergence minimization for sequence-level knowledge distillation. *arXiv preprint arXiv:2307.15190*.

Yimeng Wu, Mehdi Rezagholizadeh, Abbas Ghaddar, Md Akmal Haidar, and Ali Ghodsi. 2021. Universal-kd: Attention-based output-grounded intermediate layer knowledge distillation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7649–7661.

Minghao Wu, Abdul Waheed, Chiyu Zhang, Muhammad Abdul-Mageed, and Alham Fikri Aji. 2023. Lamini-lm: A diverse herd of distilled models from large-scale instructions. *arXiv preprint arXiv:2304.14402*.

Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. 2024. A survey on knowledge distillation of large language models. *arXiv preprint arXiv:2402.13116*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

CCL 2024

## A 实验细节

对于GPT-4反馈，我们应用如图4中的提示词模板，并将温度设置为0.7。

```
We would like to request your feedback on the performance of two AI assistants in response to the user
instruction and input displayed above. Please rate the helpfulness, relevance, accuracy, and level of detail of
their responses. Each assistant receives an overall score on a scale of 1 to 10, where a higher score indicates
better overall performance.Please first output a single line containing only two values indicating the scores
for Assistant 1 and 2 below, respectively. The two scores are separated by a space. In the subsequent line,
please provide a comprehensive explanation of your evaluation, avoiding any potential bias and ensuring
that the order in which the responses were presented does not affect your judgment.

### Instruction:
{Instruction}

### Input:
{Input}

### Assistant1:
{ground_truth}

### Assistant2:
{Model_response}
```

Figure 4: GPT4提示词模板

## B 实例分析

我们在五个实验数据集的测试集上均随机选择了不同的问答数据对作为参照（见表8,10,12,14,16），对比分析了三种仅考虑输出层蒸馏损失的基线方法与本文提出的引入了中间层蒸馏损失的蒸馏方法生成回复的内容，见表9,11,13,15,17，可以发现多数情况下本文所提出的蒸馏方法得到的蒸馏模型生成内容在语义上更接近教师模型给出的回答，且回复内容更完整、连贯、正确性更高（相较于正确答案），进一步验证了本文提出的中间层蒸馏算法对生成式大语言模型蒸馏效果的提升与在多种任务多个数据集上的泛化能力。下面是具体生成内容展示与分析：

- **DollyEval数据集**：根据表8,9分析可得，引入本文提出的中间层蒸馏算法后生成的回复（KD+OurMethod、SeqKD+OurMethod、MiniLLM+OurMethod）都比较精准而完整地回答了提问中四个国家的关联，与教师模型（Teacher）给出的回复更相似，且相比于仅计算输出层损失的基线方法（KD、SeqKD、MiniLLM）语义内容更加完整、更贴合题意。
- **SelfInst数据集**：根据表10,11分析可得，我们提出的蒸馏方法回复内容更符合原问题的答案，且相较于基线方法语义内容更丰富、具有多样性。
- **VicunaEval数据集**：根据表12,13分析可得，在面对需要逻辑推理的计算题类型问答中，我们提出的蒸馏方法更趋向于去推理算式计算的过程，而原基线方法的回答就相较之下略显杂乱、缺乏逻辑。
- **S-NI数据集**：根据表14,15分析可得，在面对摘要类型问答时，我们提出的蒸馏方法回复更精炼，且与原基线方法比较下来更正确地捕获了需提炼句子的真实含义。
- **UnNI数据集**：根据表16,17分析可得，在不考虑回复正确性的前提下，我们提出的方法基本上会以“Here...that have the most anagrams：”为开头展开回复，相较与原基线方法而言对问题的任务要求理解比较正确，回复格式也更清晰。

第二十三届中国计算语言学大会论文集，第910页-第928页，太原，中国，2024年7月25日至28日。
卷1：主会论文
(c) 2024 中国中文信息学会计算语言学专业委员会

Table 8: 实例一，DollyEval数据集指令-回复对

| | |
|---|---|
| DataSet | DollyEval |
| Instruction | What links Brazil, Uruguay, Mozambique and Angola? |
| Input | - |
| ground truth | Colonies of Portugal |

Table 9: 实例一各蒸馏方法生成回复。加粗表示的是指引入了我们提出的中间层蒸馏方法。

| Model | Response |
|---|---|
| Teacher | Brasil, Uruguay, Mozambican and Angola are all countries in South America. |
| SFT w/o KD | African lion hunting. |
| KD | Regions of the World. |
| **KD+OurMethod** | **All of them are famous Portuguese islanders.** |
| SeqKD | Brazil and Uruguay |
| **SeqKD+OurMethod** | **These 5 countries are all part of the South America.** |
| MiniLLM | The African continent is made up of countries that share a border.<br>Brazil and Mozambique are on the Atlantic and African sea and Uruguay<br>and Angola are on the Pacific and South American sea. |
| **MiniLLM+OurMethod** | **Brazil, Uruguay and Angola form the African Free Trade Area (AFFTA).** |

Table 10: 实例二，SelfInst数据集指令-回复对

| | |
|---|---|
| DataSet | SelfInst |
| Instruction | What sound does this make? |
| Input | 好 |
| ground truth | hǎo |

Table 11: 实例二各蒸馏方法生成回复。

| Model | Response |
|---|---|
| Teacher | This does not make any sound. |
| SFT w/o KD | This makes a sound when the strings are pulled together. |
| KD | This sound is a mixture of two other sounds: a truck horn and a working<br>noise machine. |
| **KD+OurMethod** | **This makes a lot of sense. It makes the listening experience more enjoyable. When you are driving on a windy road, you hear the woodwinds and also the small wailing of babies. The sound reflects off the wind and makes you want to get out of<br>the car faster.** |
| SeqKD | This makes a close approximation of the sound of a car horn. |
| **SeqKD+OurMethod** | **This makes a sound similar to a dog's bark.** |
| MiniLLM | Aah |
| **MiniLLM+OurMethod** | **Hai** |

CCL 2024


(c) 2024 中国中文信息学会计算语言学专业委员会

Table 12: 实例三，VicunaEval数据集指令-回复对

| | |
|---|---|
| DataSet | VicunaEval |
| Instruction | Given that f(x) = 5x^3 - 2x + 3, find the value of f(2). |
| Input | - |
| ground truth | To find the value of f(2), we simply substitute 2 in place of x in the given equation of f(x):<br>f(2) = 5(2)^3 - 2(2) + 3<br>f(2) = 5(8) - 4 + 3<br>f(2) = 40 - 1<br>Therefore, the value of f(2) is 39. |

Table 13: 实例三各蒸馏方法生成回复。

| Model | Response |
|---|---|
| Teacher | The value of f(2) is 7. |
| SFT w/o KD | The value of f(x) is 2. The reason it can be found in two ways is that the square of both those numbers is 4 with 5 as a reminder, which is 4 with x as a capital, so the value of x is 5. |
| KD | 2 * f = 5x^3 |
| **KD+OurMethod** | **Given that fx is 4, we can subtract sum of 4's value from the number 2 and get the result of four's value. Therefore, f(x) is 4.** |
| SeqKD | 1a = 2x/c, so 5x = 5x |
| **SeqKD+OurMethod** | **find the value of f(x -2, x -3, a -1) = 5x^4** |
| MiniLLM | F(x)=5x^3 - 2 |
| **MiniLLM+OurMethod** | **f(2) = 4.** |

Table 14: 实例四，S-NI数据集指令-回复对

| | |
|---|---|
| DataSet | S-NI |
| Instruction | In this task, you are given a text of the article.<br>Your task is to generate a headline (title) for this article. |
| Input | americans have to wait probably another year before they can relish succulent indian mangoes . |
| ground truth | time not ripe yet for indian mangoes to hit us |

CCL 2024

Table 15: 实例四各蒸馏方法生成回复。

| Model | Response |
|---|---|
| Teacher | Hurray! An indian mango is finally here!  The wait is over. Americans now have exactly one year to enjoy these juicy and juicy mangoes before they go extinct. |
| SFT w/o KD | The article will talk about the newly released indian mangoes and their pleasant aroma. |
| KD | There are a lot of positives and negative aspects to American food culture. On one hand, everyone is starving and needs food. On the other hand, American food is tasteless and does not inspire any kind of food-like expression. The best way to understand this is through the food culture. Let's call it the BIG. The Big is an expression used to describe the large number of calories eaten by Americans during the summer. So this article is about the summer calorie explosion.The phrase translated to the Big is: "BOMB". |
| **KD+OurMethod** | **The article you are provided with is "India, USA and S.Africa take centre stage in food and grooming industry". Therefore, you need to prepare a title for this article.** |
| SeqKD | The next year Indians will relish succulent indians. The reason is that bitter taste and smell of mangoes generate an indigestion in the stomach which causes a retention of the taste. Also, once the mangoes ferment long enough, they start to turn into a liquor. There is a law of evolution, indian mangoes and grapevine juice will homogenize. |
| **SeqKD+OurMethod** | **americans have been waiting for decades to relish succulent indians mangoes.** |
| MiniLLM | Americans have to wait probably another the year before they can enjoy succulent indian mangos. Much like shooting fish in a barrel. Or intelligence. |
| **MiniLLM+OurMethod** | **probably another year before they can enjoy succulent indian mango...** |

Table 16: 实例五，UnNI数据集指令-回复对

| DataSet | UnNI |
|---|---|
| Instruction | You will be given a set of words, and your task is to find the word that has the most anagrams. |
| Input | Words: ['lives', 'debit card', 'cried', 'shoulder', 'impatient']. |
| ground truth | The word with the most anagrams is "lives." |

COLING 2024

Table 17: 实例五各蒸馏方法生成回复。

| Model | Response |
|---|---|
| Teacher | 'lives', 'debt card', 'cried' |
| SFT w/o KD | 'lives', 'debits', 'cried', "shoulder", and "impatient" are all anagrams.<br>'Lives', 'cried', and'shoulder" are not anagrams. |
| KD | words: 'lives', 'cried', 'shirt','shirt','shirt', 'cried','shank', 'caught',<br>'caught', "Vote'.You will be able to find the word that best fits in the<br>given set of words. |
| **KD+OurMethod** | **Here are the words that have the most anagrams:- 'lives', 'Debit card', 'crying','shoulder', and 'impatient'.** |
| SeqKD | Here are the words that have the most anagrams: |
| **SeqKD+OurMethod** | **Here is a list of words that have the most anagrams: ['lives', ['debit card', 'broken', 'impatient'];The word 'cried' has an an anagrams: 'cried', 'broken', 'cried'.Words that do not have anagrams include: 'crying','shoulder', and 'impatient'.** |
| MiniLLM | The most anagrams can be found are: "lives", "debit card", "cried",<br>"shoulder". |
| **MiniLLM+OurMethod** | **Popular anagrams for 'lives' include 'prest', 'happy','slaughtered', 'alive'. Anagrams for 'debit card' include'stayed','somewhat', 'purchased', 'watched', 'attracted', 'behaved'. Anagrams for all others include 'dead', 'different', 'error', 'found', 'never', 'forgot', 'decide', 'likes', 'unaware','sees','more'.** |

CCL 2024

# 面向小规模大语言模型推理优化的推理路径排序方法

**李俊**[1]**, 白宇**[*1,2]**, 刘雨婷**[1]
[1]沈阳航空航天大学计算机学院，辽宁沈阳，110136
[2]多语言协同翻译技术国家地方联合工程实验室，辽宁沈阳，110136
lijun1@stu.sau.edu.cn; baiyu@sau.edu.cn; liuyuting4@stu.sau.edu.cn

## 摘要

尽管大语言模型(LLM)在自然语言处理领域取得巨大成功，但是伴随其千亿级参数规模的训练也产生了巨大的计算成本。小规模大语言模型(SLLM)作为低资源场景下实现LLM部署的可替代方案，任务处理能力与LLM尚存在明显差距。尽管上下文学习(ICL)等提示方法在一定程度上提升了SLLM的问题处理能力，但基于人工构建的提示往往需要参与者具备特定的专业领域知识，这给LLM的普适推广带来了挑战。针对以上问题，本文提出了一个基于SLLM的问题推理框架，通过在推理路径生成和答案生成两个阶段之间引入基于逐步语义验证器(SSVRP)的推理路径排序选择机制，在无人干预情况下实现SLLM推理能力提升。实验结果表明，SSVRP有效地增强了SLLM的推理性能，在4个推理任务中的平均准确率分别达到了54.3%，90.6%，64.3%和63.7%，并在其中3个推理任务中都取得了最新的SOTA结果。

**关键词：** 逐步语义验证器；小规模大语言模型；上下文学习；推理路径排序

# A Reasoning Paths Ranking Method for Reasoning Optimization of Small-scale Large Language Models

**Jun Li**[1]**, Yu Bai**[*1,2]**, Yuting Liu**[1]
[1]School of Computer Science, Shenyang Aerospace University, Shenyang, Liaoning, 110136
[2]National & Local Joint Engineering Laboratory of Multi-Language Collaborative Translation Technology, Shenyang, Liaoning, 110136
lijun1@stu.sau.edu.cn; baiyu@sau.edu.cn; liuyuting4@stu.sau.edu.cn

## Abstract

Although Large Language Models (LLM) have achieved great success in the field of natural language processing, their training with hundreds of billions of parameters also incurs huge computational costs. As an alternative to LLM deployment in low-resource scenarios, Small-scale Large Language Models (SLLM) still have a significant gap in task processing capabilities compared to LLM. Although prompting methods such as In-Context Learning (ICL) have improved the problem-handling capabilities of SLLM to a certain extent, manually constructed prompts often require participants to have specific professional domain knowledge, which poses a challenge to the universal promotion of LLM. To address the above issues, this paper proposes a question reasoning framework based on SLLM. By introducing a reasoning path ranking selection mechanism based on Step-Semantic Verifier on Reasoning Paths (SSVRP) between the two stages of reasoning path generation and answering generation, the reasoning capability

* 通讯作者 Corresponding author.


(c) 2024 中国中文信息学会计算语言学专业委员会

of SLLM is improved without human intervention. Experimental results show that SSVRP effectively enhances the reasoning performance of SLLM, with average accuracy rates of 54.3%, 90.6%, 64.3% and 63.7% in four reasoning tasks, respectively, and achieves the latest SOTA results in three of the reasoning tasks.

**Keywords:** Step-Semantic Verifier on Reasoning Paths , Small-scale Large Language Models , In-Context Learning , Reasoning Paths Ranking

## 1 引言

大型语言模型(Large Language Models, LLM)，如OpenAI的ChatGPT(OpenAI, 2022)、谷歌的Bard(Manyika and Hsiao, 2023)以及Meta的LLaMA(Touvron et al., 2023)等，在推理问答(Robinson et al., 2023)、语言翻译(Yang et al., 2023b)、情感分析(Simmering and Huoviala, 2023)和文本生成(Lu et al., 2023)等自然语言处理相关的多种任务中取得了巨大的成功，对人工智能领域的发展产生了深远的影响。

LLM通常具有复杂的模型结构以及千亿级的参数规模，这使得在本地的生产环境中部署运行LLM需要大量的计算资源成本(Barla, 2023)。以4-Bit量化的LLaMA模型为例，高效执行LLaMA-7B模型至少需要配备6GB VRAM的GPU，而要运行LLaMA-65B或70B模型，则至少需要40GB VRAM的GPU(Witt, 2024)。表1详细列出了不同参数规模的LLaMA模型对GPU资源的需求和最低部署成本。

| LLaMA模型系列 | 最小VRAM要求 | 价格(美元) |
|---|---|---|
| LLaMA/Llama-2 7B | 6GB | 280 |
| LLaMA/Llama-2 13B | 10GB | 300 |
| LLaMA/Llama-2 33B | 20GB | 2800 |
| LLaMA/Llama-2 65B/70B | 40GB | 5600 |

表 1. 4-Bit量化的LLaMA模型对GPU的需求和最低部署成本(Witt, 2024)

近期的工作(Ho et al., 2022; Magister et al., 2022)表明，小规模大语言模型(Samll-scale Large Language Model, SLLM)相对于更复杂的模型版本具有若干优势。与LLM相比，SLLM通常具有更少的参数、更小的尺寸、更快的训练和推理速度，需要更少的内存和存储，并且消耗更少的能量。这些优势可以直接为在本地部署此类模型的组织节省成本。然而，值得注意的是，SLLM的参数量的降低可能导致其总体上与LLM相比，问题处理能力相对较低。例如，大多数SLLM，即使经过微调，也难以在处理文摘任务方面超越零样本LLM(Fu et al., 2024)。

上下文学习(In-Context Learning, ICL)技术的核心在于利用一系列的输入输出对作为示例，通过给LLM提供有限的示例来指导LLM完成特定任务。该技术使得模型能够在无需参数调整的情况下预测出准确的输出。尽管ICL提示同样成功地增强了SLLM在推理任务中的能力，但提示的构建通常需要人工参与，并需要参与者具备特定的专业领域知识，这给SLLM的普适推广带来了挑战。

在应对复杂的数学或逻辑推理问题时，将一个问题拆分成一系列的中间步骤，通常能帮助LLM找到最终答案。Wei(2022)等人提出思维链(Chain-of-Thought, CoT)，并通过向LLM提供一些解题的示例，证实了CoT提示能够促使LLM将复杂的解题过程拆分为一系列的解题步骤，从而逐步引导LLM生成问题的正确答案。CoT提示作为ICL提示中的一种主流提示方法，在推理任务中表现出良好的效果。

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

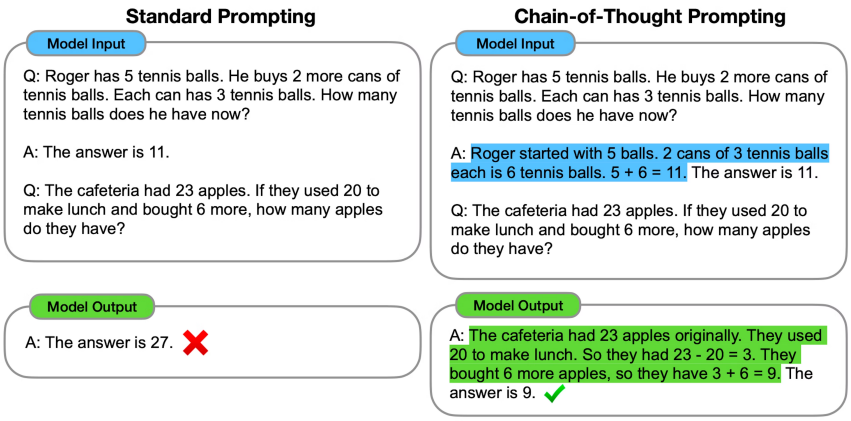



图 1. 标准的输入-输出提示和CoT提示(Wei et al., 2022)

在实际应用中，与使用*"Let's think step by step."*提示的Zero-Shot-CoT方法(Kojima et al., 2022)相比，通过人工手动构建的Manual-CoT(Wei et al., 2022)显示出更优越的性能。然而，这种高性能依赖于精心制作的示例，而创建这些示例依赖人工。Zhang等人(2022)提出Auto-CoT范式，通过自动化方式构建包含问题及其推理链的示例，旨在克服手动构建示例的限制。Wang等人(2023)提出了Self-Consistency，通过在多个推理路径中投票选择最一致的答案来提高解答的准确性，这种方法在多项算术和常识推理任务中取得了显著效果。然而，这种投票机制在面对极其多样化的推理路径时可能无法奏效。为了令LLM进行有效的推理，相关研究对不同的推理路径生成和选择进行了探索。Li等人(2023)认为，错误推理路径中的每一个步骤并非全无价值，因此提出利用DIVERSE验证器对LLM生成的各种推理路径上的每一步进行评估，以挑选出有助于LLM推理的最佳推理路径。

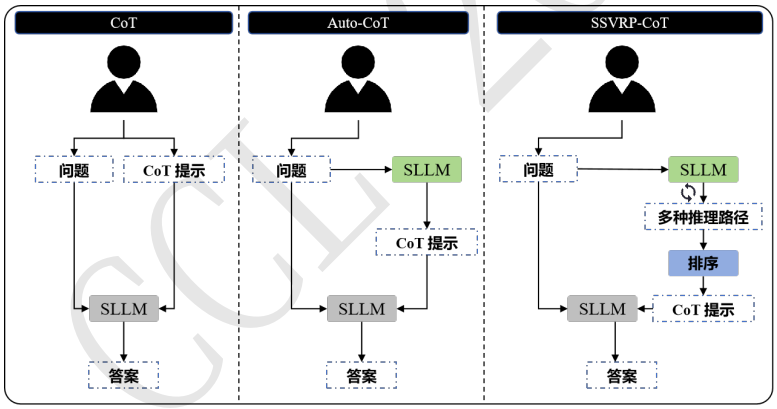



图 2. 本文方法与其它Auto-CoT方法

在本研究中，我们探索在无需人工构建任何提示的情况下，以最小的成本增强SLLM来模拟LLM在推理任务中的性能。本文的主要贡献如下：

（1）提出了一个新的框架，该框架与CoT和Auto-CoT相比，在两个SLLM之间引入基于逐步语义验证器(Step-Semantic Verifier on Reasoning Paths, SSVRP)的排序模型（如图2所示），SSVRP实现了推理路径的选择优化，从而增强SLLM在推理任务中的性能。

（2）提出了一种基于对比学习的逐步语义验证器训练方法，所有训练数据均由SLLM生成。在4个公开的推理任务数据集上进行的实验表明，答案生成的平均准确率分别达到了54.3%，90.6%，64.3%和63.7%，并在其中3个推理任务中都取得了最新的SOTA结果，达到了与基于LLM的Auto-CoT方法相当的性能。

（3）讨论了SLLM在推理路径生成和答案生成阶段的性能。我们发现不同的SLLM在推理路径生成阶段存在显著差异，而在答案生成阶段，当提供相同的推理路径时，所有模型的表现几乎相同。这表明推理路径的质量与答案的质量直接相关，表明对推理路径进行优化能够带来


(c) 2024 中国中文信息学会计算语言学专业委员会

更准确的答案。

## 2 相关工作

Wei等人(2022)提出了一种CoT提示方法，以解决LLM在理解复杂问题方面的局限性。他们手动构建CoT提示示例，逐步引导LLM将复杂问题分解为子问题，从而显著提高LLM的表现。Fu等人(2023)提出了一种根据所需推理的复杂性来选择提示的创新方法。他们优先考虑需要更复杂推理步骤的提示，从而促进详细的CoT流程。为了增强CoT提示，一些研究将输入-输出关系分解为增量推理阶段。他们采用了上下文学习(ICL)，让LLM在不同的推理阶段接受不同的提示。Self-Ask(Press et al., 2022)技术允许LLM生成与初始输入相关的补充查询，允许LLM自行查询和响应，将这些查询和响应集成到CoT提示中。iCAP(Wang et al., 2022)引入了一个动态的上下文感知提示系统，可以调整每个推理阶段的上下文。此外，Least-to-Most(Zhou et al., 2022)提示方法涉及两个阶段的ICL过程，首先将复杂问题简化为子问题，然后按顺序解决这些子问题，同时将前期阶段的答案融入到正在进行的上下文中。

Chung等人(2022a)观察到，使用包含中间推理步骤的数据训练SLLM可以显著提高SLLM的性能。STaR(Zelikman et al., 2022)可以使其通过自我生成的原因进行自我改进，从而提高SLLM在推理任务中的性能。SpecialFT(Fu et al., 2023)使用LLM作为教师模型，并利用知识蒸馏将推理能力从LLM转移到SLLM。Orca(Mukherjee et al., 2023)学习模仿LLM的推理过程，包括解释性痕迹、逐步的思维过程和其他复杂的指令。这些方法的共同特点是调整SLLM的参数并向LLM学习。

## 3 逐步语义验证器

借鉴LLM的语言理解和生成能力，我们使用了两个SLLM。如图3所示，其中SLLM-I的主要功能是生成问题的推理路径，将问题及一个初始提示*"Let's think step by step."*作为SLLM-I的输入，通过多次迭代使SLLM-I生成多条不同的候选推理路径。将利用语义验证器选出的最佳推理路径作为SLLM-II的输入，生成问题的最终答案。

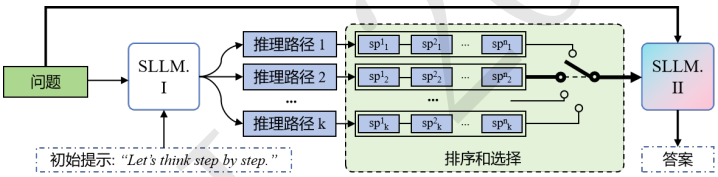



图 3. 基于推理路径排序的问题推理

算法1概述了从一组候选推理路径中选择最优推理路径的过程。

### 3.1 推理路径的排序和选择

对推理路径进行排序和选择是构建CoT提示的一个关键步骤，这一过程对于SLLM在逻辑推理和问题解决中至关重要。对推理路径进行排序和选择可以有效防止SLLM探索低质量或不相关的推理路径，从而提高SLLM在推理过程中的效率。

对于给定的问题$\mathbf{q}$，使用"."将其拆分为多条信息，记作$\mathbf{q} = [q_1, ..., q_p, ..., q_n]$，其中，$q_p$表示问题$\mathbf{q}$中的第$p$条信息。假设SLLM-I为问题$\mathbf{q}$生成$k$条推理路径，使用"."将每一条推理路径其拆分为多个推理步骤，构成序列$\mathbf{R} = \left[sp_1^{(1)}, ..., sp_{m_1}^{(1)}, ..., sp_j^{(i)}, ..., sp_1^{(k)}, ..., sp_{m_k}^{(k)}\right]$，$sp_j^{(i)}$表示第$i$条推理路径中的第$j$个推理步骤，$i = 1, ..., k$，$j = 1, ..., m_i$。在这里，最优的推理路径是指能够引导SLLM-II生成正确答案的推理路径。考虑到在实际场景中，SLLM-II能否利用推理路径得出正确答案具有不确定性，我们把与问题提供的语义信息最接近的推理路径作为最佳推理路径，其在序列$\mathbf{R}$中的位置表示为$Pos^*$。

$$Pos^* = \underset{k}{argmax} \left[ \sum_{p=1}^{n} \sum_{j=1}^{m_k} sim\left(q_p, sp_j^{(k)}\right) / m_k \right], \tag{1}$$

其中，函数$sim(\cdot)$的作用是预测问题$q$中第$p$条信息和第$k$条候选推理路径中第$j$个推理步骤之间语义相似性。逐步语义验证器的训练和用于最佳推理路径排序和选择的过程如图4所示。


(c) 2024 中国中文信息学会计算语言学专业委员会

### 3.2 逐步语义验证器

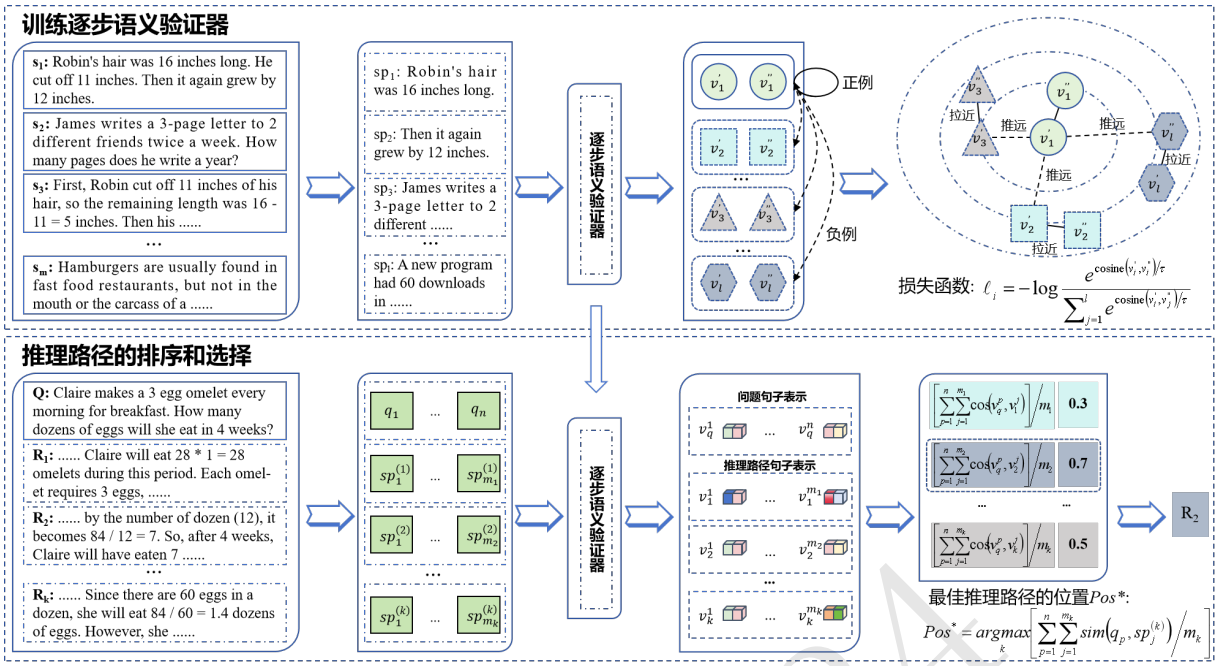



图 4. 逐步语义验证器的训练和用于最佳推理路径排序和选择的过程

为了实现函数$sim(\cdot)$，我们提出了一个逐步语义验证器（SSVRP），SSVRP将问题中的每一条信息和候选推理路径中的每一个推理步骤进行编码，使其映射到潜在语义空间。令$v_q^p$表示问题$q$的第$p$个信息编码后的表示向量，$v_i^j$表示第$i$条推理路径的第$j$个推理步骤编码后的表示向量，通过计算$v_q^p$和$v_i^j$之间夹角的余弦值实现函数$sim(\cdot)$的输出。

为了学习推理路径中每一个推理步骤的语义特征，在不使用LLM和人工干预的情况下，利用SLLM创建用于训练逐步语义验证器的数据集。为了保证推理路径的多样性，我们采用迭代的方式利用多个不同的SLLM为训练问题集中的每一个问题生成多条推理路径。在这个推理路径的生成过程中，使用了初始化提示*"[Question] + Let's think step by step."*。

逐步语义验证器的训练遵循SimCSE(Gao et al., 2021)中无监督的对比学习框架，并将问题本身视为一条推理路径，并入训练数据集。我们将该训练数据集表示为句子集合$\{sp_j\}_{j=1}^{l}$，采用dropout噪声和$sp_j^+ = sp_j$作为数据增强的一种形式。将同一个句子两次经过dropout操作，得到句子的两个不同表示向量互为正例，不同句子的嵌入表示互为负例。

$$\mathbf{v}_j^z = f_\theta(sp_j, z), \tag{2}$$

其中$z$是dropout的随机掩码（默认$p$=0.1），我们只需将相同的输入以两次提供给编码器并获得两个不同的嵌入$v'$, $v''$。在这项工作中，我们使用的对比学习损失函数与SimCSE相同：

$$\ell_i = -\log \frac{e^{cosine\left(v_i', v_i''\right)/\tau}}{\sum_{j=1}^{l} e^{cosine\left(v_i', v_j''\right)/\tau}}, \tag{3}$$

其中$\tau$是温度超参数，$cosine(v', v'')$是余弦相似度$\frac{v'^{\top} v''}{\|v'\|\|v''\|}$。我们使用目标函数$\ell_i$对预训练语言模型*DeBERTa-base*(He et al., 2021)进行微调，并利用微调后的模型对输入句子进行编码，得到问题中的每一条信息和候选推理路径中的每一个推理步骤在潜在语义空间中的向量表示。


(c) 2024 中国中文信息学会计算语言学专业委员会

**算法 1** 推理路径的排序和选择

**Input:** A question: **q**, Number of Reason paths generated by $SLLM\text{-}I$: $k$.
**Output:** The Optimal Position in the List of Reasoning Paths $Pos^*$.

Initial the prompt: $P \leftarrow$ "Let's think step by step."
Initial the Optimal position: $Pos^* \leftarrow 0$
Initial the Optimal value: $Score^* \leftarrow$ -1
$S[] \leftarrow [Q, SLLM\text{-}I(Q, P, k)]$ //$SLLM$-$I$生成问题的推理路径
**q**[] $\leftarrow$ Split $Q$ by "."
**for** each reasoning path $R_i$ in $S/Q$ **do**
  $R_i[] \leftarrow$ Split $R_i$ by "."
  $Score_i \leftarrow 0$
  **for** each sentence $q_p$ in **q** **do**
    $Score_{q_p} \leftarrow 0$
    $v_q^p \leftarrow Semantic\_Verifier(q_p)$
    //计算此个推理路径中所有推理步骤与其中一个问题信息的语义相似度的和
    **for** each sentence $sp_j^{(i)}$ in $R_i$ **do**
      $n \leftarrow 0$
      $v_i^j \leftarrow Semantic\_Verifier(sp_j^{(i)})$
      $Score_{sp_j} \leftarrow cosine(v_q^p, v_i^j)$
      $Score_{q_p} \leftarrow Score_{q_p} + Score_{sp_j}$
    **end for**
    $Score_i \leftarrow Score_{q_i} + Score_i$
  **end for**
  $Score_i \leftarrow Score_i / n$
  //找出最佳推理路径的位置
  **if** $Score_i > Score^*$ **then**
    $Score^* \leftarrow Score_i$
    $Pos^* \leftarrow i$
  **end if**
**end for**
**return** $Pos^*$

## 4 实验

### 4.1 实验步骤

#### 4.1.1 数据集描述

本文实验采用4个公开数据集测试SLLM在算术推理和常识推理任务上的表现。其中，GSM8K(Cobbe et al., 2021)、MultiArith(Roy and Roth, 2015)以及SVAMP(Patel et al., 2021)3个数据集主要用来评估SLLM解决复杂的多步算术问题的能力，CommonsenseQA(Talmor et al., 2019)数据集用来评估SLLM解决常识推理问题的能力。实验数据集的详细信息如表2所示。

#### 4.1.2 训练数据

本文使用的训练数据集包含8593个算术推理问题和9741个常识推理问题。为了确保生成的推理路径的多样性，我们利用3个不同的SLLM（LLaMA2、Baichuan2和ChatGLM3）为每个问题生成9个推理路径，每个SLLM生成3个推理路径。由此，共生成77257个算术推理问题的推理路径和87669个常识推理问题的推理路径。采用分隔符"."将推理路径句子拆分成多个推理步骤，并通过随机采样构造算术推理问题推理步骤集合和常识推理问题推理步骤集合。

#### 4.1.3 对比模型

对比模型采用4个基于LLM的增强模型和9个基于SLLM的增强模型。基于LLM的增强模型包括分别使用Greedy Decode和Self-Consistency提示增强的GPT-3(davinci)、分别使

CCL 2024

| 数据集 | 问题类型 | 训练数据的规模 | 测试数据的规模 |
|---|---|---|---|
| GSM8K | 数学应用题 | 7473 | 1319 |
| MultiArith | 数学应用题 | 420 | 180 |
| SVAMP | 数学应用题 | 700 | 300 |
| CommonsenseQA | 常识单选题 | 9741 | 1221 |

表 2. 实验数据的详细信息

用Greedy Decode和Self-Consistency提示增强的GPT-3(text-davinci-002)。基于SLLM的增强模型包括分别使用Greedy Decode、Self-Consistency和DIVERSE提示增强的FlanT5-XL(3B)，使用Greedy Decode、Self-Consistency和DIVERSE提示增强的LLaMA2(7B)，使用Greedy Decode、Self-Consistency和DIVERSE提示增强的Baichuan2(7B)，使用Greedy Decode、Self-Consistency和DIVERSE提示增强的ChatGLM3(6B)。这些比较模型的组件描述如下：

• **LLM**

**GPT-3**(Brown et al., 2020)拥有1750亿的参数，是OpenAI开发的开源大型语言模型。它以高精度和复杂性执行各种自然语言处理任务。在本文中，我们引用了Li等人(2023)报告的两个公共引擎davinci和code-davinci-002的实验结果。

• **SLLM**

**FlanT5-XL(3B)**(Chung et al., 2022b)是T5模型的一种，XL版本的FlanT5模型具有30亿的参数量。

**LLaMA2(7B)**(Touvron et al., 2023)是一个经过预训练和微调的大型语言模型，具有70亿的参数。它是Meta开发的LLaMA2的最小版本。

**Baichuan2(7B)**(Yang et al., 2023a)是一个包含70亿的参数的大规模多语言的语言模型，是百川智能开发的Baichuan2的最小版本。

**ChatGLM3(6B)**(Zeng et al., 2022)是一个轻量级开源预训练对话模型，包含60亿的参数。它是ChatGLM3的最小版本，ChatGLM3是中英文训练的双语大语言模型，属于智普AI开发的大型语言模型系列。

• **提示增强方法**

**Greedy Decode**(Kojima et al., 2022)是一种将问题与初始提示 *"Let's think step by step."* 一起输入LLM的方法。这种方法鼓励LLM逐步解决问题，从而得出最终答案。

**Self-Consistency**(Wang et al., 2023)使用Few-Shot-CoT提示生成多个推理路径和对应的答案，然后选择频率最高的答案作为最终答案。

**DIVERSE**(Li et al., 2023)是最近研究中最先进的方法，可增强LLM的推理能力。此方法需要LLM(code-davinci-002)生成的推理路径的数据集来训练验证器。

#### 4.1.4 超参数设置

在SLLM-I生成候选的推理路径时，期望让SLLM-I生成的推理路径具有多样性，因此开启了SLLM-I的随机采样模式。在SLLM-II回答问题时，期望让SLLM-II生成更一致的内容，因此在此阶段关闭了SLLM-II的随机采样模式。在训练阶段，我们引入了多头自注意力机制，设置num_heads的数量为8，其它超参数设置如表3所示。

### 4.2 实验结果

本文采用准确率(Accuracy)作为基于推理路径排序的问题推理框架在算术推理和常识推理任务中的性能评价指标，实验结果如表4所示。

### 4.3 实验分析

#### 4.3.1 逐步语义验证器的有效性

为了验证基于SSVRP进行推理路径选择的有效性，将SSVRP与推理路径的随机选择方法进行比较。此外，为了验证将推理路径分割为多个推理步骤的必要性，将采用推理步骤进行语义验证的SSVRP与采用推理路径进行语义验证的SVRP方法进行比较。


(c) 2024 中国中文信息学会计算语言学专业委员会

| 模型 | 超参数 | 取值 | | | |
|---|---|---|---|---|---|
| SLLM | do_sample | SLLM-I | True | SLLM-II | False |
| | top_k | | 10 | | 1 |
| | temperature | | 0.7 | | 0 |
| | k | | 3 | | - |
| DeBERTa | dropout | 0.1 | | | |
| | num_heads | 8 | | | |
| | $\tau$ | 0.05 | | | |
| | learning_rate | 2e-5 | | | |
| | batch_size | 16 | | | |
| | max_sequence_length | 256 | | | |

表 3. SLLM和DeBERTa的超参数设置

| Method | GSM8K | MultiArith | SVAMP | CommonsenseQA | Average |
|---|---|---|---|---|---|
| GPT-3 (davinci, 175B): | | | | | |
| Greedy Decode | 8.7 | 31.4 | 21.2 | 48.2 | 27.4 |
| Self-Consistency | 18.9 | 68.6 | 44.6 | 57.4 | 47.4 |
| GPT-3 (text-davinci-002, 175B): | | | | | |
| Greedy Decode | 37.1 | 70.7 | 60.0 | 65.5 | 58.3 |
| Self-Consistency | 58.2 | 88.4 | 78.2 | 72.9 | 74.4 |
| FlanT5-XL (3B): | | | | | |
| Greedy Decode | 5.4 | 10.0 | 8.7 | 77.8 | 25.5 |
| Self-Consistency | 2.9 | 6.1 | 7.3 | **81.9** | 24.6 |
| DIVERSE (LLM) | **6.4** | 14.4 | 11.3 | 79.3 | 27.9 |
| SSVRP (SLLM) | 6.2 | **16.7** | **14.0** | 80.6 | **29.4** |
| LLaMA2 (7B): | | | | | |
| Greedy Decode | 24.5 | 68.3 | 39.7 | 43.7 | 44.1 |
| Self-Consistency | 22.9 | 68.9 | 41.3 | 43.8 | 44.2 |
| DIVERSE (LLM) | 22.8 | **69.7** | **44.3** | 45.6 | 45.5 |
| SSVRP (SLLM) | **25.9** | 66.1 | 42.0 | **48.7** | **45.7** |
| Baichuan2 (7B): | | | | | |
| Greedy Decode | 30.9 | 74.4 | 48.7 | 55.4 | 52.4 |
| Self-Consistency | 29.0 | 79.4 | 47.3 | 57.4 | 53.3 |
| DIVERSE (LLM) | **34.1** | 80.6 | **54.7** | 58.5 | 56.9 |
| SSVRP (SLLM) | 33.5 | **82.6** | 53.7 | **59.1** | **57.2** |
| ChatGLM3 (6B): | | | | | |
| Greedy Decode | 52.9 | 86.7 | 59.3 | 60.8 | 64.9 |
| Self-Consistency | 45.8 | 93.9 | 60.0 | 61.4 | 65.3 |
| DIVERSE (LLM) | 52.7 | **94.6** | 62.3 | 62.7 | 68.1 |
| SSVRP (SLLM) | **54.3** | 90.6 | **64.3** | **63.7** | **68.2** |

表 4. SLLM在不同的推理任务中的有效性比较。最好结果用黑体表示，第二好的结果用下划线表示。

SVRP方法与SSVRP方法的主要区别在于，SVRP方法不对推理路径按步骤进行拆分。具体来说，SLLM-I模型会为每一道题生成多个推理路径，每一条推理路径都是一个包含多个解题步骤的解题过程。我们将题目信息和完整的推理路径作为训练语义验证器的训练数据。

如图5所示的实验结果表明，与随机选择推理路径的方法相比，在ChatGLM3(6B)上，SSVRP在多个推理任务中的性能均优于随机方法和SVRP方法，说明SSVRP的有效性。

#### 4.3.2 推理路径的增强效果

实验结果表明，与Greedy Decode、Self-Consistency和DIVERSE方法相比，在LLaMA2(7-


(c) 2024 中国中文信息学会计算语言学专业委员会

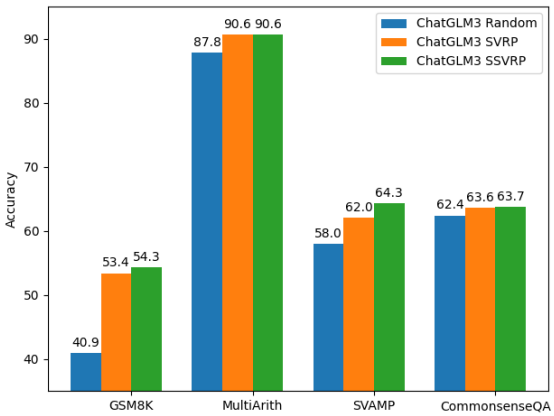



图 5. 在ChatGLM3上的推理路径选择方法的比较

B)上，SSVRP在推理任务中的平均准确率分别提高了3.5%，3.3%和0.4%，在Baichuan2(7-B)上，分别提高了8.4%，6.8%和0.5%，在ChatGLM3(6B)上，分别提高了4.8%，4.3%和0.2%。此外，在更小参数量的FlanT5-XL(3B)上，SSVRP在推理任务中的平均准确率分别提高了13.3%，16.3%和5.1%。表明SSVRP方法对于100亿参数以下的SLLM是有效的。与不同提示增强方法(Greedy Decode和Self-Consistency)在davinci和text-davinci-002上的比较结果表明，SSVRP增强的SLLM能够达到LLM的推理效果。

为了进一步研究SLLM-I生成的推理路径的质量对SLLM-II推理能力的影响，我们进行了SLLM-I和SLLM-II的组合实验，实验结果如表5所示。

| Combination | GSM8K | MultiArith | SVAMP | CommonsenseQA |
|---|---|---|---|---|
| LLaMA2: | | | | |
| +LLaMA2 | 25.9 | 66.1 | 42.0 | 48.7 |
| +Baichuan2 | 26.8 | 68.9 | 43.0 | 57.0 |
| +ChatGLM3 | **27.1** | **70.6** | **45.3** | **58.6** |
| Baichuan2: | | | | |
| +LLaMA2 | 32.7 | 76.1 | 44.0 | 46.7 |
| +Baichuan2 | 33.5 | 82.6 | 53.7 | 59.1 |
| +ChatGLM3 | **34.4** | **83.9** | **53.7** | **59.5** |
| ChatGLM3: | | | | |
| +LLaMA2 | 53.7 | 88.9 | 61.3 | 54.5 |
| +Baichuan2 | 53.7 | 88.9 | 63.3 | 60.9 |
| +ChatGLM3 | **54.3** | **90.6** | **64.3** | **63.7** |

表 5. 不同数据集中组合模型的有效性对比，下划线表示SLLM-I，+表示SLLM-II

实验结果表明，当提供相同的推理路径时，各个模型(SLLM-II)在答案生成阶段的表现几乎相同。这也表明，推理路径的优劣是影响问题推理结果的主要因素，有效的推理路径排序与选择方法有助于提升SLLM的推理能力。

## 5 结论

本文提出了一个基于SLLM的问题推理框架，通过在推理路径生成和答案生成两个阶段之间引入基于逐步语义验证器(SSVRP)的推理路径排序选择机制，在无人干预情况下实现SLLM推理能力提升。相关实验结果表明，推理路径的优劣直接影响SLLM的推理结果，有效的推理路径排序与选择方法有助于提升SLLM的推理能力。利用多个（两个）SLLM的协作方法能够使它们获得与单个LLM相当的推理能力。语义验证器能够通过选择最优推理路径来增


(c) 2024 中国中文信息学会计算语言学专业委员会

强SLLM的推理能力，在推理路径评价过程中，采用推理步骤进行语义验证的SSVRP方法优于采用推理路径进行语义验证的SVRP方法。

## 致谢

本文受到国家自然科学基金(U1908216)的资助。

## 参考文献

Nilesh Barla. 2023. Deploying large nlp models: Infrastructure cost optimization. Accessed on April 8th, 2024.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022a. Scaling instruction-finetuned language models. *arXiv e-prints*, pages arXiv–2210.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022b. Scaling instruction-finetuned language models.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Yao Fu, Hao Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. 2023. Specializing smaller language models towards multi-step reasoning. *arXiv preprint arXiv:2301.12726*.

Xue-Yong Fu, Md Tahmid Rahman Laskar, Elena Khasanova, Cheng Chen, and TN ShashiBhushan. 2024. Tiny titans: Can smaller large language models punch above their weight in the real world for meeting summarization? *ArXiv*, abs/2402.00841.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2022. Large language models are reasoning teachers. *arXiv e-prints*, pages arXiv–2212.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. 2023. Making language models better reasoners with step-aware verifier. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5315–5333, Toronto, Canada, July. Association for Computational Linguistics.

Albert Lu, Hongxin Zhang, Yanzhe Zhang, Xuezhi Wang, and Diyi Yang. 2023. Bounding the capabilities of large language models in open text generation with prompt constraints. In Andreas Vlachos and Isabelle Augenstein, editors, *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1982–2008, Dubrovnik, Croatia, May. Association for Computational Linguistics.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2022. Teaching small language models to reason. *arXiv preprint arXiv:2212.08410*.


(c) 2024 中国中文信息学会计算语言学专业委员会

James Manyika and Sissie Hsiao. 2023. An overview of bard: an early experiment with generative ai. *AI. Google Static Documents*, 2.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4. *arXiv preprint arXiv:2306.02707*.

OpenAI. 2022. Chatgpt: A large-scale generative model for open-domain chat. https://chat.openai.com/.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are nlp models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2080–2094.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. 2022. Measuring and narrowing the compositionality gap in language models. *ArXiv*, abs/2210.03350.

Raquel B Robinson, Karin Johansson, James Collin Fey, Elena Márquez Segura, Jon Back, Annika Waern, Sarah Lynne Bowman, and Katherine Isbister. 2023. Edu-larp@ chi. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–5.

Subhro Roy and Dan Roth. 2015. Solving general arithmetic word problems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1743–1752.

Paul F Simmering and Paavo Huoviala. 2023. Large language models for aspect-based sentiment analysis. *arXiv preprint arXiv:2310.18025*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Boshi Wang, Xiang Deng, and Huan Sun. 2022. Iteratively prompt pre-trained language models for chain of thought. In *Conference on Empirical Methods in Natural Language Processing*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Allan Witt. 2024. Computer hardware required to run llama ai model locally (gpu, cpu, ram, ssd). Accessed on April 8th, 2024.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023a. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Wen Yang, Chong Li, Jiajun Zhang, and Chengqing Zong. 2023b. Bigtrans: Augmenting large language models with multilingual translation capability over 100 languages. *arXiv preprint arXiv:2305.18098*.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alexander J. Smola. 2022. Automatic chain of thought prompting in large language models. *ArXiv*, abs/2210.03493.

CCL 2024

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V Le, et al. 2022. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations*.

CCL 2024

# 面向中文实体识别的Transformers模型句子级非对抗鲁棒性研究

**王立帮　王裴岩*　沈思嘉**
沈阳航空航天大学计算机学院，辽宁沈阳 110136
wangpy@sau.edu.cn
{wanglibang, shensijia}@stu.sau.edu.cn

## 摘要

基于Transformers的中文实体识别模型在标准实体识别基准测试中取得了卓越性能，其鲁棒性研究也受到了广泛关注。当前，中文实体识别模型在实际部署中所面临的句子级非对抗鲁棒性问题研究不足，该文针对该问题开展了研究。首先，该文从理论上分析并发现了Transformer中自注意力、相对位置嵌入及绝对位置嵌入对模型鲁棒性的负面影响。之后，提出了实体标签增强和滑动窗口约束的鲁棒性增强方法，并从理论上证明了提出方法能够提升Transformers模型的实体识别鲁棒性。最后，通过在3个中文数据集的实验，研究了4种基于Transformer的实体识别模型的脆弱性，所提出方法使模型的鲁棒性F1值提升最高可达4.95%。

**关键词：** 实体识别；非对抗鲁棒性；Transformers模型；句子级扰动

# On Sentence-level Non-adversarial Robustness of Chinese Named Entity Recognition with Transformers Model

**Libang Wang　Peiyan Wang　Sijia Shen**
School of Computer Science, Shenyang Aerospace University,
Shenyang, Liaoning 110136, China
wangpy@sau.edu.cn
{wanglibang, shensijia}@stu.sau.edu.cn

## Abstract

Chinese entity recognition models based on Transformers have shown exceptional performance in standard benchmarks, with robustness studies in this area gaining extensive attention. However, research on sentence-level non-adversarial robustness issues faced by Chinese entity recognition models in practical deployment is insufficient. This study addresses this gap by first theoretically analyzing the negative impacts of the Transformer model's self-attention, relative position embeddings, and absolute position embeddings on model robustness. It then proposes robustness enhancement methods through entity label augmentation and sliding window constraints, theoretically proving these methods can improve the Transformers model's entity recognition robustness. Experimental results on three Chinese datasets reveal vulnerabilities in four Transformer-based entity recognition models, with the proposed methods increasing model robustness F1 scores by up to 4.95%.

* 通讯作者
©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版
**基金项目：**辽宁省应用基础研究(2022JH2/101300248)，国家自然科学基金资助项目(U1908216)

**Keywords:** entity recognition , non-adversarial robustness , transformers model , sentence-level perturbation

## 1 引言

命名实体识别（Named Entity Recognition，NER）指从文本中识别出具有特定意义的实体，并将其分类为预定义类别，如人名、地名、组织和机构等。其是自然语言处理（Natural Language Processing，NLP）领域中的一项基础任务，在信息检索 (Tan et al., 2021)、问答系统 (Min et al., 2021)、对话系统 (Wang et al., 2020)等多个领域中具有广泛应用。

目前，以Transformer为典型代表的神经网络架构模型在NLP领域的多项任务中均取得了卓越性能 (Vaswani et al., 2017; Kenton and Toutanova, 2019)，但此类模型在应对扰动输入时的脆弱性也引发了研究者对其安全性的思考。最近一些研究 (Liang et al., 2018; Wu et al., 2023; Zhu et al., 2023)表明，此类模型容易受到扰动示例影响。通过扰动输入序列来干扰模型，最终使其预测结果发生改变。

NER作为研究和工业应用中最常见的NLP任务之一，近年来随着深度学习兴起，其也取得了长足发展。特别是，Transformer架构的提出为此带来了新动力。借助Transformer强大的长序列建模与并行计算能力，NER模型在识别性能和运行效率上都取得了显著提升 (Zhang and Yang, 2018; Li et al., 2020)。同时，基于深层Transformer构建的预训练模型结合微调或提示范式的广泛应用也为NER研究带来新突破 (Hu et al., 2022)。

NER作为NLP领域中的一个基础任务，其对构建更智能、更精准的NLP系统具有关键作用，因此NER模型的鲁棒性问题也引起了研究者的关注。现有研究 (Srinivasan and Vajjala, 2023)证明NER模型对输入数据的微小变化很敏感。表 1展示了一个扰动实例，在所列出的扰动序列输出中，模型预测输出的“深圳特区”应被标记为地缘政治（GPE）实体。然而，正如扰动后的输出所示，其却被标记为组织（ORG）实体。显然，一些看似无害的微小扰动正在改变模型的预测结果。

| 类型 | 实例 |
|---|---|
| 原始序列1 | 香港$_{GPE}$ 回归后，深$_{GPE}$ 港$_{GPE}$ 两地经贸关系得到加强，深$_{GPE}$ 港$_{GPE}$ 西部$_{LOC}$ 通道等三大基础设施衔接项目目前进展顺利。 |
| 预测输出 | 香港$_{GPE}$ 回归后，深港$_{GPE}$ 两地经贸关系得到加强，深港$_{GPE}$ 西部通道等三大基础设施衔接项目目前进展顺利。 |
| 原始序列2 | 深圳$_{GPE}$ 特区$_{LOC}$ 今年推进国有企业改革成效显著，全市今年工业总产值预计达一千二百五十五亿元，比去年增长百分之十七点五。 |
| 预测输出 | 深圳特区$_{GPE}$ 今年推进国有企业改革成效显著，全市今年工业总产值预计达一千二百五十五亿元，比去年增长百分之十七点五。 |
| 扰动序列 | 原始序列1 + 原始序列2 |
| 预测输出 | 香港$_{GPE}$ 回归后，深港$_{GPE}$ 两地经贸关系得到加强，深港$_{GPE}$ 西部通道等三大基础设施衔接项目目前进展顺利。深圳特区$_{ORG}$ 今年推进国有企业改革成效显著，全市今年工业总产值预计达一千二百五十五亿元，比去年增长百分之十七点五。 |

Table 1: 对OntoNote4.0中序列进行扰动后BERT-CRF模型预测实例

目前，对于NER模型鲁棒性的研究多集中在英文环境下 (Namysl et al., 2020; Lin et al., 2021; Das and Paik, 2022)，其通过多种方法构造不同扰动样本对模型鲁棒性展开评估，其中多数方法也取得了很高的攻击成功率，最新研究将此推广到了多种语言环境下 (Srinivasan and Vajjala, 2023)。但Wang等人 (2022a)表明，现有模型鲁棒性研究中所构造的扰动样本多是非自然的，这些样本在现实场景中并不会出现，同时这类样本大多是不流畅的。因此，根据这些非自然扰动样本来判断模型鲁棒性是不合理的。


(c) 2024 中国中文信息学会计算语言学专业委员会

基于上述分析，在中文NER任务中，结合目前中文NER模型在实际部署中所面临的句子级自然扰动现象，本文开展对中文NER模型在句子级自然扰动下鲁棒性的研究。同时本文遵循计算机视觉中Drenkow等人 (2021)对模型鲁棒性的进一步划分，将这类在句子级自然扰动下模型鲁棒性称作模型的句子级非对抗鲁棒性，同时提出两个主要问题：

- 中文NER下基于Transformer构建的模型是否会受到句子级自然扰动样本的影响?
- 如果此类NER模型在该类自然扰动影响下不鲁棒，应如何提高模型句子级非对抗鲁棒性?

针对问题1，本文遵循NER模型在实际应用中所出现的句子级自然扰动场景，在多个中文NER数据集上构建句子级鲁棒性评估数据集以评估模型鲁棒性；针对问题2，本文引入两种约束策略来改善NER模型的句子级非对抗鲁棒性。在这项工作中，本文通过一系列实验对中文NER模型进行鲁棒性评估，最终得到如下发现：

- 与LSTM相比，使用Transformer或预训练语言模型结合微调或提示学习所构建的NER模型更容易受到句子级自然扰动影响。
- Transformer中的位置嵌入和自注意力模块削弱了句子级自然扰动下模型鲁棒性。其中绝对位置嵌入模块对模型鲁棒性的影响独立于自注意力模块，而相对位置嵌入模块则依赖自注意力模块，只有自注意力模块发挥作用时其才会对模型鲁棒性产生影响。
- 在训练阶段增强模型对已知实体的关注度可提高模型的句子级非对抗鲁棒性，但这种策略提升效果有限。
- 在模型中引入滑动窗口注意力机制可以改善句子级自然扰动下模型的鲁棒性。

## 2 相关工作

### 2.1 模型微调

BERT (Kenton and Toutanova, 2019)首次证明了在大规模预训练语言模型的基础上，通过微调可以在各种任务上实现卓越的性能。通过在大规模通用语料库上进行无监督预训练，然后在特定任务或领域的有监督微调中逐步调整模型参数，以实现更高级别的自然语言理解和生成，从而为各种文本处理任务提供强大的通用基础。当前，由大语料训练的预训练模型在中文序列标注方面表现出色，使用BERT字符特征的模型在中文NER和中文词性标注等方面的表现也远远优于基于静态嵌入的方法 (Hu et al., 2022)。

### 2.2 提示学习

提示学习是一种有监督学习方法，旨在减少对庞大标注数据集的依赖。与预训练加微调范式不同，提示学习将下游任务转换为与模型预训练任务更一致的形式。基于提示学习的模型通过应用提示功能将输入序列转化为具有未填充槽的各种提示模板，通过填充这些槽，可以完成各种NLP任务，例如推荐系统 (Wang et al., 2022b)、情感分析 (Li et al., 2023)、事件抽取 (Ma et al., 2022b)等。同时，该方法的发展也使NER下零样本和少样本学习成为可能 (Ma et al., 2022a)。本文主要贡献在于探讨了在受到句子级自然扰动时，基于提示学习的中文NER模型鲁棒性所受影响，扩展了当前的研究。

### 2.3 模型鲁棒性

鲁棒性是深度学习模型的一个重要属性，其是指模型在面对与训练数据不同的输入时依然能够产生预期输出的能力。在计算机视觉中，Drenkow等人 (2021)对模型鲁棒性做了进一步划分，包括对抗鲁棒性和非对抗鲁棒性，其中非对抗鲁棒性也被称为自然鲁棒性。对抗鲁棒性衡量了模型对经过故意破坏的输入数据进行准确分类的能力，其目标是通过欺骗模型使其以高置信度输出错误的预测结果。这种类型的鲁棒性取决于攻击者对模型和数据属性的了解程度。

NLP下模型对抗鲁棒性研究通常采用多种形式对模型进行干扰，以改变模型的输出结果。Zhu等人 (2023)为评估大语言模型（Large Language Models，LLMs）对对抗性提示的鲁棒性，推出了一个系统基准，称之为PromptBench，旨在评估、理解和分析LLMs的鲁棒性。在多个数据集上的实验结果显示，不同类型的攻击在效果上存在显著差异。其中单词级攻击威胁最大，可致所有数据集的平均性能下降33%，字符级攻击次之，句子级攻击的威胁最低。Wang等人 (2023)对ChatGPT在NLP任务中的对抗鲁棒性和分布外泛化能力进行了深入评估，结果表明，以ChatGPT为代表的LLMs在鲁棒性和分布外泛化能力上较以往模型有显著提升，但与人类水平相比，仍有一定差距。


(c) 2024 中国中文信息学会计算语言学专业委员会

NER下模型对抗鲁棒性研究涵盖了对抗性攻击的多个粒度，包括字符级、单词级和句子级攻击。Namysl等人 (2020)、Lin等人 (2021)、Das等人 (2022)及Srinivasan等人 (2023)在文本中随机翻转输入序列中的字符或单词、随机单词序列或者添加分散模型注意力的句子来混淆模型，使模型预测出错。尽管对抗性鲁棒性可以在模型面对恶意攻击时保持模型的稳健性，但非对抗性鲁棒性能够在自然引发的数据扰动场景下保持模型性能和预测结果一致性。因此，非对抗鲁棒性可被视为一个理想的模型属性，在模型实际部署中，自然扰动下的数据最为常见。

## 3 实体识别模型的鲁棒性评估

### 3.1 问题定义

本文将NER下的句子级鲁棒性问题定义为模型在受到句子级扰动时所出现的预测结果不一致现象。即对任意原始序列$\mathbf{X}^o = \{c_i\}_{i=1}^p$，使用自然扰动样本生成算法$A$可构造其扰动序列$\mathbf{X}^r = A(\mathbf{X}^o)$。该扰动序列中某个字符输入已训练好的模型，模型将以较高置信度输出与原始标签不同的结果，如式(1)所示：

$$c^o \in \mathbf{X}^o,\ c^r \in \mathbf{X}^r;\ \exists c^o = c^r,\ f(c^o) \neq f(c^r) \tag{1}$$

同时，本文将模型在受到句子级自然扰动时的扰动成功率$Asr$定义为：

$$Asr = \frac{1}{l}\sum_i \mathbb{I}(f(c_i^o) \neq f(c_i^r)) \tag{2}$$

式中，$l$为序列$\mathbf{X}^o$的长度，$\mathbb{I}(x)$为指示函数，$c_i^o$和$c_i^r$分别为原始序列$\mathbf{X}^o$和扰动序列$\mathbf{X}^r$中索引为$i$处的字符。

对任意一段输入序列$\mathbf{X} = \{c_i\}_{i=1}^n$，各字符经过词嵌入后的输入表征如式(3)所示：

$$\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n = Embedding(c_1, c_2, \cdots, c_n) \tag{3}$$

式中，$Embedding()$可为静态词嵌入如Glove (Pennington et al., 2014)等，亦或动态词嵌入如BERT (Kenton and Toutanova, 2019)等。

在句子级自然扰动下，Transformer模型中输入向量$\boldsymbol{x}$在经过自注意力模块后的输出向量$\boldsymbol{t}$被改变，最终干扰了模型的预测结果。假设原始序列$\mathbf{X}^o$和扰动序列$\mathbf{X}^r$中字符$\boldsymbol{c}$在经过自注意力模块后的输出向量分别为$\boldsymbol{t}^o$和$\boldsymbol{t}^r$，若$\cos(\boldsymbol{t}^o, \boldsymbol{t}^r)$低于某个阈值时会改变模型对于当前字符的预测结果，可列出相应表达式如式(4)所示：

$$\cos(\boldsymbol{t}^o, \boldsymbol{t}^r) = \frac{\boldsymbol{t}^o \cdot \boldsymbol{t}^r}{\|\boldsymbol{t}^o\| \|\boldsymbol{t}^r\|} < 1 - \varepsilon \tag{4}$$

式中，$\varepsilon$用于衡量扰动的大小。

此时，结合式(4)可将式(2)中原始扰动成功率$Asr$调整为式(5)来衡量Transformer中在扰动样本生成算法$A$影响下的扰动成功率：

$$Asr = \frac{1}{l}\sum_i \mathbb{I}(\cos(\boldsymbol{t}_i^o, \boldsymbol{t}_i^r) < 1 - \varepsilon) \tag{5}$$

### 3.2 鲁棒性评估数据集的构建

遵循AdvGLUE (Wang et al., 2021)中设计的扰动样本生成原则，并结合中文NER模型在实际应用场景中最常见的句子级扰动情况，本文使用基于分心的句子级扰动方法，扰动标准数据集中的原始序列来生成鲁棒性评估数据集。具体来说，通过拼接测试集中相邻两段序列$\mathbf{X}^{o_1}$和$\mathbf{X}^{o_2}$来生成句子级扰动序列$\mathbf{X}^r$，即$\mathbf{X}^r = A(\mathbf{X}^{o_1}) = [\mathbf{X}^{o_1}; \mathbf{X}^{o_2}]$，该扰动不会改变原始序列的语义。


(c) 2024 中国中文信息学会计算语言学专业委员会

### 3.3 评价指标

针对模型性能和鲁棒性的评估，本文在多种数据集上评估了实体级微平均F1（Micro-F1）分数，以比较不同模型的性能。目前NER任务下模型鲁棒性评估主要报告了两种F1变体的实验结果，以量化模型在标准数据集和鲁棒性评估数据集上的性能，分别是标准F1和扰动F1。

- **标准F1 (Standard-F1，S-F1)**：在标准数据集上评估模型所得的F1分数，其综合了模型的精确性和完整性，这是NER任务中使用最广泛的指标。
- **扰动F1 (Perturbation-F1，P-F1)**：在所构造的鲁棒性评估数据集上评估模型所得的F1分数。P-F1越大表明模型在鲁棒性评估数据集上的预测准确性越高。

在评估NER模型鲁棒性时，上述两种评价指标存在局限。如表 1所展示的模型鲁棒性评估实例，对于原始序列1和2的输出结果和对应的扰动输入序列，若仅计算其S-F1和P-F1，不难看出最终F1值是相等的，也即为此时在这类自然扰动影响下模型是鲁棒的。然而，仔细观察这三段序列各自预测结果，可以看出原始序列2中“深圳特区”的预测标签和扰动序列中“深圳特区”的预测标签是不一致的，即在这类自然扰动影响下模型仍是不鲁棒的。

鉴于此，本文参考Huang等人 (2021)使用标准精度、扰动精度、鲁棒精度来评估语义解析器的性能，在当前任务下，本文额外报告了一种F1变体的实验结果，称之为鲁棒F1。

- **鲁棒F1 (Robust-F1，R-F1)**：综合标准和鲁棒性评估两个数据集下模型的评估结果，以标准数据集下对应字符的预测标签作为鲁棒性评估数据集下该字符的真实标签，来评估模型在扰动数据集下的F1分数。R-F1越大表明模型在标准数据集和鲁棒性评估数据集上的预测结果的差异性越小，也即为模型在这两个数据集上预测结果的一致性越高。

### 3.4 自注意力和位置嵌入模块影响分析

Transformer作为一种基于自注意力机制的神经网络模型，其能够对序列中的每个字符进行全局建模，并在各字符间建立联系。与循环神经网络模型相比，Transformer具有更好的并行性能和更短的训练时间。Transformer的核心之一是自注意力机制，其作用是为输入序列中各字符间计算一个权重，然后将这些加权后的输入向量作为输出。

如下针对NER任务对Transformer架构作简单介绍，本文仅讨论其中的编码器部分。其由自注意力层和前馈网络层（Feed-Forward Networks，FFN）组成，每个子层后都有残差连接和层归一化，其中FFN是具有非线性变换的多层感知器。Transformer通过H个注意力头对序列$\mathbf{X}$进行自注意力，然后拼接H个头的输出得到最终结果。为简化计算过程，如下计算将忽略自注意力模块中注意力头数的影响，同时假设所有原始序列的长度都为$l$。在自注意力模块中，可以使用三个不同的参数矩阵$\mathbf{W}_q$、$\mathbf{W}_k$、$\mathbf{W}_v$将$\boldsymbol{x}_i$映射为三个新向量$\boldsymbol{q}_i = \boldsymbol{x}_i\mathbf{W}_q$、$\boldsymbol{k}_i = \boldsymbol{x}_i\mathbf{W}_k$、$\boldsymbol{v}_i = \boldsymbol{x}_i\mathbf{W}_v$。此时，经过自注意模块后输出向量$\boldsymbol{t}_i$计算式如式(6)所示：

$$
\begin{aligned}
\hat{\alpha}_{i,j} &= \frac{\boldsymbol{q}_i \boldsymbol{k}_j^{\mathrm{T}}}{\sqrt{d}} \\
\hat{\boldsymbol{\alpha}}_i &= [\hat{\alpha}_{i,1},\ \hat{\alpha}_{i,2},\ ...,\ \hat{\alpha}_{i,l}] \\
\mathbf{A} &= [\hat{\boldsymbol{\alpha}}_1,\ \hat{\boldsymbol{\alpha}}_2,\ ...,\ \hat{\boldsymbol{\alpha}}_l]^{\mathrm{T}} \\
\alpha_{i,j} &= \mathrm{S}\left(\mathbf{A}_i\right)_j \\
\boldsymbol{t}_i &= \sum_j \alpha_{i,j} \boldsymbol{v}_j
\end{aligned}
\tag{6}
$$

式中，$\mathrm{S}\left(\boldsymbol{x}\right)$为对$\boldsymbol{x}$的Softmax运算。

Transformer中除了自注意力模块，还同时存在位置嵌入模块。常用的位置嵌入方式有两种，分别为绝对位置嵌入和相对位置嵌入。对于使用绝对位置嵌入的神经网络模型，例如Transformer，其输入向量$\boldsymbol{x}$主要由两部分构成，分别为标识嵌入向量$\boldsymbol{x}_e$和绝对位置嵌入向量$\boldsymbol{x}_p$，即$\boldsymbol{x} = \boldsymbol{x}_e + \boldsymbol{x}_p$。对于使用相对位置嵌入的神经网络模型，例如TENER (Yan et al., 2019)，其相对位置嵌入计算如式(7)所示：

$$
\hat{\alpha}_{i,j} = \boldsymbol{q}_i \boldsymbol{k}_j^{\mathrm{T}} + \boldsymbol{q}_i \boldsymbol{r}_{i-j}^{\mathrm{T}} + \boldsymbol{u} \boldsymbol{k}_j^{\mathrm{T}} + \boldsymbol{v} \boldsymbol{r}_{i-j}^{\mathrm{T}} \tag{7}
$$

式中，$\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^{d_k}$是可学习的参数，$\boldsymbol{r}_{i-j} \in \mathbb{R}^{d_k}$是相对位置嵌入。


(c) 2024 中国中文信息学会计算语言学专业委员会

### 3.4.1 自注意力对模型鲁棒性的影响

在句子级自然扰动下，Transformer模型的全局自注意力权重分布改变，导致模型不鲁棒。若假设当前模型中未使用位置嵌入模块，则此时输入向量$\boldsymbol{x}$满足$\cos(\boldsymbol{x}^o, \boldsymbol{x}^r) = 1$。在句子级自然扰动影响下，受扰动序列与原始序列相比其对应注意力权重分布被改变，导致输出向量发生改变。此时，可列出经过自注意力模块后的输出向量$\boldsymbol{t}$计算如式(8)所示：

$$\begin{aligned}\boldsymbol{t}_i^o &= \begin{cases}\sum\limits_{j\in C_1} \mathrm{S}\left([\hat{\alpha}_{i,1}, \hat{\alpha}_{i,2}, ..., \hat{\alpha}_{i,l}]\right)_j \boldsymbol{v}_j, & i \in C_1 \\ \sum\limits_{j\in C_2} \mathrm{S}\left([\hat{\alpha}_{i,l+1}, \hat{\alpha}_{i,l+2}, ..., \hat{\alpha}_{i,2l}]\right)_j \boldsymbol{v}_j, & i \in C_2\end{cases} \\ \boldsymbol{t}_i^r &= \sum_{j\in C_1\cup C_2} \mathrm{S}\left([\hat{\alpha}_{i,1}, \hat{\alpha}_{i,2}, ..., \hat{\alpha}_{i,2l}]\right)_j \boldsymbol{v}_j, \ i \in C_1 \cup C_2\end{aligned} \tag{8}$$

式中，$C_1 = C_{\mathbb{N}}[1, l]$，$C_2 = C_{\mathbb{N}}[l+1, 2l]$。

由式(8)可以看出，由于自注意力权重分布改变，导致满足式(5)的输出向量$\boldsymbol{t}$数量增加，提高了扰动成功率，最终影响了模型的鲁棒预测。

### 3.4.2 绝对位置嵌入对模型鲁棒性的影响

为便于后续计算$\cos(\boldsymbol{t}^o, \boldsymbol{t}^r)$，本文按照扰动序列构建方式将原始序列$\mathbf{X}^{o_1}$和$\mathbf{X}^{o_2}$中各字符对齐到扰动序列$\mathbf{X}^r$中对应位置字符。即在$i < l$时，$\boldsymbol{x}_i^o$赋值为$\mathbf{X}^{o_1}$中索引为$i$处字符的表征向量$\boldsymbol{x}_i^{o_1}$，而在$i > l$时，$\boldsymbol{x}_i^o$赋值为$\mathbf{X}^{o_2}$中索引为$i-l$处字符的表征向量$\boldsymbol{x}_{i-l}^{o_2}$。对于使用绝对位置嵌入的模型，输入向量$\boldsymbol{x}$计算如式(9)所示：

$$\begin{aligned}\boldsymbol{x}_i^o &= \begin{cases}\boldsymbol{x}_{i:e}^{o_1} + \boldsymbol{x}_{i:p}^{o_1}, & i \in C_1 \\ \boldsymbol{x}_{i-l:e}^{o_2} + \boldsymbol{x}_{i-l:p}^{o_2}, & i \in C_2\end{cases} \\ \boldsymbol{x}_i^r &= \begin{cases}\boldsymbol{x}_i^o, & i \in C_1 \\ \boldsymbol{x}_{i-l:e}^{o_2} + \boldsymbol{x}_{i:p}^r, & i \in C_2\end{cases}\end{aligned} \tag{9}$$

为排除自注意力对模型鲁棒性的影响，在扰动序列中，式(6)中矩阵$\mathbf{A}$加入额外约束以排除自注意力模块的干扰，计算公式如式(10)所示：

$$\begin{aligned}\mathbf{W} &= [-\infty] \in \mathbb{R}^{l\times l} \\ \mathbf{A} &= [\hat{\boldsymbol{\alpha}}_1, \hat{\boldsymbol{\alpha}}_2, ..., \hat{\boldsymbol{\alpha}}_{2l}]^{\mathrm{T}} + \begin{bmatrix}\mathbf{O} & \mathbf{W} \\ \mathbf{W} & \mathbf{O}\end{bmatrix}\end{aligned} \tag{10}$$

式中，$\mathbf{O}$为零矩阵。

此时，经过自注意力模块后的输出向量$\boldsymbol{t}$计算如式(11)所示：

$$\begin{aligned}\boldsymbol{t}_i^o &= \begin{cases}\sum\limits_{j\in C_1} \mathrm{S}\left([\hat{\alpha}_{i,1}, \hat{\alpha}_{i,2}, ..., \hat{\alpha}_{i,l}]\right)_j \boldsymbol{x}_j^o \mathbf{W}_v, & i \in C_1 \\ \sum\limits_{j\in C_2} \mathrm{S}\left([\hat{\alpha}_{i,l+1}, \hat{\alpha}_{i,l+2}, ..., \hat{\alpha}_{i,2l}]\right)_j \boldsymbol{x}_j^o \mathbf{W}_v, & i \in C_2\end{cases} \\ \boldsymbol{t}_i^r &= \begin{cases}\boldsymbol{t}_i^o, & i \in C_1 \\ \sum\limits_{j\in C_2} \mathrm{S}\left([\hat{\alpha}_{i,l+1}, \hat{\alpha}_{i,l+2}, ..., \hat{\alpha}_{i,2l}]\right)_j \boldsymbol{x}_j^r \mathbf{W}_v, & i \in C_2\end{cases}\end{aligned} \tag{11}$$

观察式(11)可以看出，当$i \in C_1$时，此时模型的预测结果是鲁棒的，当$i \in C_2$时，模型可能会出现预测结果的改变，此时模型是不鲁棒的，综合上述两种情况如式(12)所示：

$$\cos(\boldsymbol{t}_i^o, \boldsymbol{t}_i^r) = \begin{cases}1, & i \in C_1 \\ p < 1, & i \in C_2\end{cases} \tag{12}$$

因此，和未使用绝对位置嵌入的模型相比，绝对位置嵌入的使用会降低模型在句子级自然扰动下的鲁棒性。同时，绝对位置嵌入对模型鲁棒性的负面影响不受限于自注意力模块。

CCL 2024

### 3.4.3 相对位置嵌入对模型鲁棒性的影响

对于使用相对位置嵌入的模型，式(7)所示给出了相对位置嵌入计算过程。此时，输入向量$\boldsymbol{x}$的计算如式(13)所示：

$$\boldsymbol{x}_i^o = \boldsymbol{x}_i^r = \begin{cases} \boldsymbol{x}_{i:e}^{o_1}, & i \in C_1 \\ \boldsymbol{x}_{i-l:e}^{o_2}, & i \in C_2 \end{cases} \tag{13}$$

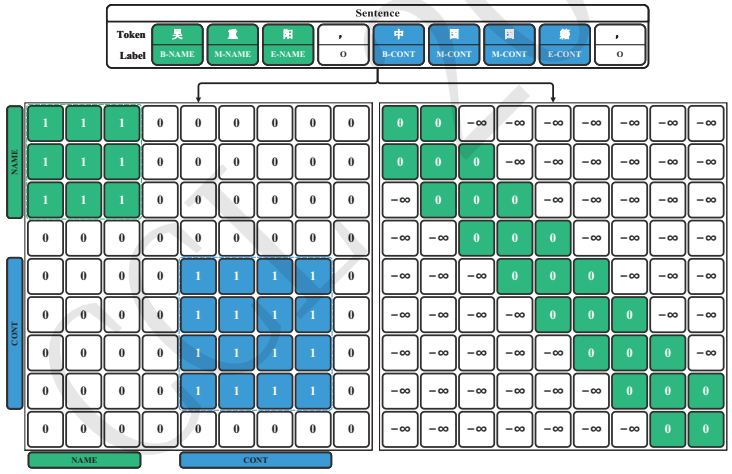
这里为排除自注意力模块的干扰依旧沿用3.4.2节中式(10)所示的约束条件。同时，由于没有绝对位置嵌入向量$\boldsymbol{x}_{i:p}$，原始序列$\mathbf{X}^o$和扰动序列$\mathbf{X}^r$中对应位置处字符的向量表征是相等的。这时经过自注意力模块后的输出向量$\boldsymbol{t}$计算如式(14)所示：

$$\boldsymbol{t}_i^o = \boldsymbol{t}_i^r = \begin{cases} \sum\limits_{j \in C_1} \mathrm{S}\left([\hat{\alpha}_{i,1},\ \hat{\alpha}_{i,2},\ ...,\ \hat{\alpha}_{i,l}]\right)_j \boldsymbol{x}_j^o \mathbf{W}_v, & i \in C_1 \\ \sum\limits_{j \in C_2} \mathrm{S}\left([\hat{\alpha}_{i,l+1},\ \hat{\alpha}_{i,l+2},\ ...,\ \hat{\alpha}_{i,2l}]\right)_j \boldsymbol{x}_j^o \mathbf{W}_v, & i \in C_2 \end{cases} \tag{14}$$

观察式(14)可以看出，在排除自注意力模块影响后，对$\forall i \in C_1 \cup C_2$，$\cos\left(\boldsymbol{t}_i^o, \boldsymbol{t}_i^r\right) \equiv 1$，此时相对位置嵌入并不会影响模型在句子级自然扰动下的鲁棒性。只有同时考虑自注意力模块对模型鲁棒性的影响时，相对位置嵌入对模型鲁棒性的影响才会显现。

## 4 提高中文实体识别模型鲁棒性方法

基于3.4节对Transformer结构的分析，本文引入两种约束策略来提高模型的句子级非对抗鲁棒性，这两种约束策略对应矩阵的构建方式如图 1所示。




(a) 实体标签增强矩阵　　(b) 滑动窗口约束矩阵

Figure 1: 两种约束策略对应矩阵的构建

### 4.1 实体标签增强

对于训练集中任意序列$\mathbf{X}$，可得到所有实体的跨度集合$S = \{s_1, s_2, ..., s_m\}$和每个实体的跨度$s$所对应标签$y$。给定起始索引$b_i$和结束索引$e_i$，跨度$s_i$可表示为$(b_i, e_i)$。如图 1所示，此段序列中可能跨度集合与标签分别为$\{(1, 3), (5, 8)\}$和$\{NAME, CONT\}$。如图 1 (a)所示，实体标签增强矩阵$\mathbf{W}_{cl}$为定义在$\mathbb{R}^{l \times l}$上的矩阵，计算方法如式(15)所示，$l$为序列$\mathbf{X}^o$的长度。

$$\begin{aligned} \mathbf{W}_{cl} &= \beta\left[a_{ij}\right] \in \mathbb{R}^{l \times l} \\ a_{ij} &= \sum_{s_n \in S} \mathbb{1}_{C_{\mathbb{N}}[b_n, e_n]}(i)\, \mathbb{1}_{C_{\mathbb{N}}[b_n, e_n]}(j) \end{aligned} \tag{15}$$


(c) 2024 中国中文信息学会计算语言学专业委员会

式中，$\mathbb{1}_A(x)$为指示函数，$\beta \in [0,1]$为超参数。若字符$c_i$和$c_j$在同一实体跨度内，则$a_{ij}$为1，否则为0。$\mathbf{W}_{cl}$描述了同实体内字符间的关联关系。

**定理1**：对受到句子级自然扰动影响的任意一段序列$\mathbf{X}^o$，有扰动序列$\mathbf{X}^r$，实体标签增强矩阵$\mathbf{W}_{cl}$，自注意力权重矩阵$\mathbf{A}$，使用$\mathbf{A}$获得$\boldsymbol{x}$对应输出$\boldsymbol{t}$，基于式(16)增强矩阵$\mathbf{A}$，可获得新输出$\tilde{\boldsymbol{t}}$，则有式(17)成立。证明过程见附录。

$$\mathbf{A} = (\mathbf{W}_{cl} + \mathbf{J}) \odot [\hat{\boldsymbol{\alpha}}_1,\ \hat{\boldsymbol{\alpha}}_2,\ ...,\ \hat{\boldsymbol{\alpha}}_l]^{\mathrm{T}} \tag{16}$$

$$\cos\left(\tilde{\boldsymbol{t}}_i^o,\ \tilde{\boldsymbol{t}}_i^r\right) > \cos\left(\boldsymbol{t}_i^o,\ \boldsymbol{t}_i^r\right) \tag{17}$$

式中，$\odot$为哈达玛积，$\mathbf{J} \in \mathbb{R}^{l \times l}$为全1矩阵。

### 4.2 滑动窗口约束

Transformer的注意力分布是按比例缩放且平滑的，但结合3.4节分析，在NER任务下，若期望模型在句子级自然扰动下更鲁棒，稀疏注意力是更适合的，并非所有字符都需要被关注。对于给定字符，几个上下文字符就足以判断其标签，平滑注意力可能包含一些嘈杂信息。

如图 1 (b)所示，滑动窗口约束矩阵$\mathbf{W}_{cs}$为定义在$\mathbb{R}^{l \times l}$上的矩阵，计算公式如式(18)所示。

$$\begin{aligned} \mathbf{W}_{cs} &= [a_{ij}] \in \mathbb{R}^{l \times l} \\ a_{ij} &= \begin{cases} -\infty, & |i-j| > w \\ 0, & otherwise \end{cases} \end{aligned} \tag{18}$$

式中，$w \in C_{\mathbb{N}}[0, l-1]$为滑动窗口注意力模式中的窗口长度。

**定理2**：对受到句子级自然扰动影响的任意一段序列$\mathbf{X}^o$，有扰动序列$\mathbf{X}^r$，滑动窗口约束矩阵$\mathbf{W}_{cs}$，自注意力权重矩阵$\mathbf{A}$，使用$\mathbf{A}$获得$\boldsymbol{x}$对应输出$\boldsymbol{t}$，基于式(19)增强，可获得新输出$\tilde{\boldsymbol{t}}$，则有式(17)成立。其中，本文的滑动窗口约束遵循Beltagy等人 (2020)所设计的滑动窗口注意力模式。证明过程见附录。

$$\mathbf{A} = [\hat{\boldsymbol{\alpha}}_1,\ \hat{\boldsymbol{\alpha}}_2,\ ...,\ \hat{\boldsymbol{\alpha}}_l]^{\mathrm{T}} + \mathbf{W}_{cs} \tag{19}$$

## 5 实验设置与实验结果分析

### 5.1 实验数据

本文在三个中文NER数据集对模型的句子级非对抗鲁棒性展开评估。

- **Weibo** (Peng and Dredze, 2015)：此数据集源自于社交媒体领域中的新浪微博。
- **Resume** (Zhang and Yang, 2018)：此数据集由上市公司高管的1027份简历经处理而得。
- **OntoNotes4.0** (Pradhan et al., 2013)：此数据集由标注一个大型语料库得到，覆盖了多种文本类型，包括新闻文章、社交媒体文本等，以及不同领域内容，如政治、经济、医学等。在本文中使用其中文部分，并使用了与Che等人 (2013)相同的预处理方法。

### 5.2 模型

本文依照3.2节构建鲁棒性评估数据集，并在下列模型上评估模型句子级非对抗鲁棒性。

- **BiLSTM-CRF** (Huang et al., 2015)：应用BiLSTM-CRF到NER任务中，使模型有效利用已有和新输入的特征。同时由于CRF层的存在，模型能够利用句子层面的标注信息。
- **Lattice-LSTM** (Zhang and Yang, 2018)：通过将潜在词信息融入到字符信息中并构建有向无环图（晶格结构），利用LSTM门控机制计算字符与词之间的相关性，以提升模型的识别性能。
- **TENER** (Yan et al., 2019)：使用基于相对位置嵌入的Transformer编码器建模句子中的字符和单词级特征，同时删除缩放点积自注意力中比例因子以提升Transformer识别性能。
- **BERT或RoBERTa-CRF**：使用BERT (Kenton and Toutanova, 2019)、RoBERTa (Liu et al., 2019)这类预训练语言模型作为编码器，同时使用CRF作为解码器，在特定领域数据集上对预训练语言模型进行微调。其中，RoBERTa是对BERT进行调整后得到的模型。这一调整涉及对BERT预训练任务的改进，以提高模型稳健性。
- **EntLM** (Ma et al., 2022a)：将NER任务表述为无需任何模板的LM问题，放弃模板构建过程，同时保留预训练模型的单词预测范式，来预测实体位置处与类别相关的标签词。

CCL 2024

### 5.3 模型鲁棒性评估实验

为评估中文NER模型的句子级非对抗鲁棒性，本文设计了一组句子级自然扰动影响实验。通过3.2节所示的句子级扰动样本生成方式在原始序列中引入句子级自然扰动，观察模型对于原始和扰动两段序列中对应命名实体的识别结果，最终实验结果如表 2所示。

| 模型 | Weibo | | | Resume | | | OntoNotes4.0 | | |
|---|---|---|---|---|---|---|---|---|---|
| | S-F1 | P-F1 | R-F1 | S-F1 | P-F1 | R-F1 | S-F1 | P-F1 | R-F1 |
| **BiLSTM-CRF** | 49.60 | 50.20 | 95.41 | 93.57 | 93.44 | 98.96 | 61.39 | 60.71 | 95.50 |
| **Lattice-LSTM** | 57.91 | 58.30 | **98.31** | 94.65 | 94.23 | **99.35** | 73.79 | 73.46 | **98.96** |
| **TENER** | 55.51 | 54.92 | 92.45 | 95.19 | 94.84 | 98.68 | 73.02 | 72.99 | 95.28 |
| **BERT-CRF** | 67.83 | 69.41 | 91.28 | 95.55 | 95.46 | 97.67 | 81.65 | 81.32 | 94.94 |
| **RoBERTa-CRF** | 67.28 | 68.31 | 89.67 | 96.10 | 95.80 | 98.00 | 81.59 | 81.68 | 95.23 |
| **EntLM** | 60.65 | 59.33 | 86.87 | 94.24 | 94.47 | 96.03 | 70.92 | 70.53 | 84.71 |

Table 2: 句子级自然扰动下不同数据集中各模型的标准、扰动、鲁棒F1值

表 2实验结果揭示了Transformers模型在处理扰动序列时的脆弱性，从表 2可以观察到：

1. 与LSTM模型相比，Transformers模型在句子级自然扰动下鲁棒性较差。本文认为这是由于Transformer中的自注意力机制允许模型对输入序列中所有位置的字符信息进行加权处理，这使得Transformer易受这类扰动影响。通过变动序列中某些词，使自注意力机制关注到不相关部分，最终干扰模型预测结果。
2. 与Transformer及预训练语言模型加微调相比，预训练语言模型加提示学习所构建模型在应对句子级自然扰动时的鲁棒性较差。
3. 与Transformer相比，预训练语言模型结合微调或提示所构建模型虽可借助预训练语言模型在大规模语料上进行上游学习优势，有助于模型更好地捕捉句子中的语义和语法信息，提高模型对实体边界和类别识别的准确性，但实验结果也显示这类模型在面对句子级自然扰动时的鲁棒性更为脆弱。这一现象与计算机视觉中Zhang等人 (2023)在ImageNet上观测结果相契合，即预训练模型在目标数据集上进行微调的方法不只会使模型获得更强的特征提取能力，也会将预训练模型带有的非鲁棒特征转移到微调后的模型中。

### 5.4 自注意力及位置嵌入对鲁棒性影响实验

为研究句子级自然扰动下基于Transformer构建的模型鲁棒性较差的主要原因，基于3.4节的分析，发现Transformer易受这类扰动影响的主要原因是由于其中的自注意模块和位置嵌入模块，因此本文进行了如下实验：依次约束Transformer下的自注意力模块和位置嵌入模块，同时观察两种模块对模型鲁棒性的影响，最终实验结果如表 3所示。其中PosE表示位置嵌入模块，Attn表示自注意力模块。

表 3展示了Transformer中自注意力模块和位置嵌入模块单独和共同作用时所对应的实验结果，从中可以观察到：

1. Transformer中位置嵌入模块和自注意力模块的共同作用削弱了句子级自然扰动下模型鲁棒性。但此操作也会使模型预测精度显著降低，尤其对EntLM这类基于预训练语言模型加提示学习范式所构建模型，其预测精度下降幅度最大。
2. 对于使用相对位置嵌入模型，如TENER，单独移除自注意力模块与同时移除位置嵌入模块和自注意力模块都可以显著提高模型鲁棒性，并且这两种方式产生的影响相同。
3. 对于使用绝对位置嵌入模型，如BERT或RoBERTa加CRF、EntLM等，自注意力模块和位置嵌入模块共同影响了模型在句子级自然扰动下的鲁棒性，仅约束其中之一无法有效改善模型在受到此类自然扰动时的鲁棒性。


(c) 2024 中国中文信息学会计算语言学专业委员会

| 模型 | PosE | Attn | Weibo | | | Resume | | | OntoNotes4.0 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | S-F1 | P-F1 | R-F1 | S-F1 | P-F1 | R-F1 | S-F1 | P-F1 | R-F1 |
| **TENER** | ✓ | ✓ | 55.51 | 54.92 | 92.45 | 95.19 | 94.84 | 98.68 | 73.02 | 72.99 | 95.28 |
| | ✓ | × | 53.44 | 53.46 | **98.13** | 92.72 | 92.37 | **99.66** | 68.27 | 68.21 | **99.29** |
| | × | ✓ | 51.10 | 50.07 | 89.15 | 94.01 | 92.90 | 97.60 | 69.47 | 68.68 | 92.63 |
| | × | × | 53.44 | 53.46 | **98.13** | 92.72 | 92.37 | **99.66** | 68.27 | 68.21 | **99.29** |
| **BERT-CRF** | ✓ | ✓ | 67.83 | 69.41 | 91.28 | 95.55 | 95.46 | 97.67 | 81.65 | 81.32 | 94.94 |
| | ✓ | × | 23.88 | 23.82 | 92.41 | 82.39 | 80.70 | 93.00 | 53.03 | 53.02 | 91.15 |
| | × | ✓ | 38.02 | 31.13 | 67.88 | 79.06 | 72.38 | 81.20 | 59.89 | 57.02 | 80.55 |
| | × | × | 22.64 | 22.64 | **100.00** | 82.84 | 82.84 | **100.00** | 54.04 | 53.69 | **99.37** |
| **RoBERTa-CRF** | ✓ | ✓ | 67.28 | 68.31 | 89.67 | 96.10 | 95.80 | 98.00 | 81.59 | 81.68 | 95.23 |
| | ✓ | × | 22.37 | 22.45 | 95.17 | 81.74 | 80.08 | 92.90 | 53.39 | 53.14 | 92.39 |
| | × | ✓ | 39.14 | 31.66 | 61.03 | 80.42 | 74.22 | 81.31 | 59.11 | 56.94 | 79.90 |
| | × | × | 22.25 | 22.25 | **100.00** | 82.56 | 82.56 | **100.00** | 54.45 | 54.10 | **99.41** |
| **EntLM** | ✓ | ✓ | 60.65 | 59.33 | 86.87 | 94.24 | 94.47 | 96.03 | 70.92 | 70.53 | 84.71 |
| | ✓ | × | 0.90 | 0.91 | 94.50 | 21.43 | 19.12 | 83.93 | 6.04 | 6.98 | 82.52 |
| | × | ✓ | 26.87 | 23.73 | 70.23 | 48.30 | 37.80 | 68.32 | 27.11 | 23.74 | 64.11 |
| | × | × | 0.89 | 0.89 | **100.00** | 14.70 | 14.70 | **100.00** | 4.35 | 4.35 | **100.00** |

Table 3: 自注意力及位置嵌入对鲁棒性影响实验下不同数据集中各模型标准、扰动、鲁棒F1值

### 5.5 模型鲁棒性增强实验

针对Transformers模型在句子级自然扰动下所表现出的不鲁棒现象，本文在模型训练与推理阶段引入多种约束策略以提高模型对于此类自然扰动的鲁棒性。不同约束策略下模型的评估结果如表 4所示，其中cl表示使用实体标签增强策略，cs表示使用滑动窗口约束策略。

| 模型 | Weibo | | | Resume | | | OntoNotes4.0 | | |
|---|---|---|---|---|---|---|---|---|---|
| | S-F1 | P-F1 | R-F1 | S-F1 | P-F1 | R-F1 | S-F1 | P-F1 | R-F1 |
| **TENER** | 55.51 | 54.92 | 92.45 | 95.19 | 94.84 | 98.68 | 73.02 | 72.99 | 95.28 |
| **+ cl** | 55.32 | 54.62 | 93.87 | 94.49 | 94.58 | 98.81 | 72.67 | 72.83 | 95.73 |
| **+ cs** | 55.73 | 55.60 | **97.40** | 94.96 | 94.47 | **99.27** | 73.25 | 73.12 | **99.42** |
| **BERT-CRF** | 67.83 | 69.41 | 91.28 | 95.55 | 95.46 | 97.67 | 81.65 | 81.32 | 94.94 |
| **+ cl** | 68.09 | 67.45 | 92.22 | 95.81 | 96.16 | **98.43** | 81.12 | 80.95 | 95.14 |
| **+ cs** | 66.27 | 66.12 | **93.41** | 95.25 | 94.69 | 98.22 | 79.96 | 79.49 | **96.28** |
| **RoBERTa-CRF** | 67.28 | 68.31 | 89.67 | 96.10 | 95.80 | 98.00 | 81.59 | 81.68 | 95.23 |
| **+ cl** | 68.85 | 67.37 | 90.57 | 95.60 | 95.17 | 97.96 | 81.56 | 81.48 | 95.21 |
| **+ cs** | 67.29 | 66.82 | **92.96** | 95.70 | 95.85 | **99.24** | 78.52 | 78.68 | **96.23** |
| **EntLM** | 60.65 | 59.33 | 86.87 | 94.24 | 94.47 | 96.03 | 70.92 | 70.53 | 84.71 |
| **+ cl** | 60.32 | 59.37 | 84.52 | 94.84 | 93.24 | 95.52 | 72.27 | 72.19 | 85.57 |
| **+ cs** | 59.64 | 59.29 | **88.57** | 93.77 | 93.49 | **98.06** | 71.89 | 71.57 | **89.63** |

Table 4: 模型鲁棒性增强实验下不同数据集中各模型的标准、扰动、鲁棒F1值

观察表 4实验结果可以得出：

1. 与原始模型相比，本文引入的两种鲁棒性增强策略可提高Transformers模型在句子级自然扰动下的鲁棒性，同时也注意到这两种鲁棒性增强策略也会对模型预测精度产生影响。


(c) 2024 中国中文信息学会计算语言学专业委员会

2. 与实体标签增强策略相比，滑动窗口约束策略所带来的鲁棒性提升更为明显。同时由于这种约束策略将Transformer中自注意力限制在了某个预先设定的范围，因此该策略会对模型预测精度产生较大的负面影响。
3. 两种约束策略在Transformer下的表现优于预训练语言模型加微调或者提示学习。本文推测，预训练语言模型中多层Transformer块的堆叠可能也会对句子级自然扰动下模型鲁棒性产生影响。未来工作中，将对预训练语言模型中多层Transformer块的堆叠情况进行研究。

## 6 总结

目前，基于Transformer构建的中文NER系统在实际应用中已取得先进性能，但在实际应用场景下发现，此类系统对输入数据的敏感性较高。当前，NER鲁棒性研究多专注于英语，同时此类研究多专注于对抗性攻击，这类攻击所构造的多数样本在现实场景中并不会出现，因此，本文开展了面向中文NER任务的Transformers模型的非对抗鲁棒性研究。同时由于句子级自然扰动在现有中文NER系统实际应用阶段更为常见，因此本文聚焦于句子级自然扰动下模型鲁棒性。最后，针对模型在句子级自然扰动下的不鲁棒现象，本文引入了两种约束策略以提高模型鲁棒性。

在多个不同规模数据集下实验结果表明：首先，基于Transformer构建的NER模型鲁棒性评估结果均低于LSTM，表明Transformer在面对句子级自然扰动时更为脆弱；其次，同时约束Transformer中的位置嵌入和自注意力模块可有效改善模型在句子级自然扰动下的鲁棒性，但这种策略也会造成模型预测精度的显著降低；最后，本文引入了两种约束策略，实验结果表明其可改善Transformers模型在句子级自然扰动下的鲁棒性，但同时也注意到这两种策略在预训练语言模型加微调或者提示学习范式下所获得鲁棒性受益较低。

## 参考文献

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Wanxiang Che, Mengqiu Wang, Christopher D Manning, and Ting Liu. 2013. Named entity recognition with bilingual constraints. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 52–62.

Sudeshna Das and Jiaul Paik. 2022. Resilience of named entity recognition models under adversarial attack. In *Proceedings of the First Workshop on Dynamic Adversarial Data Collection*, pages 1–6.

Nathan Drenkow, Numair Sani, Ilya Shpitser, and Mathias Unberath. 2021. A systematic review of robustness in deep learning for computer vision: Mind the gap? *arXiv preprint arXiv:2112.00639*.

Shulin Hu, Huajun Zhang, Xuesong Hu, and Jinfu Du. 2022. Chinese named entity recognition based on bert-crf model. In *2022 IEEE/ACIS 22nd International Conference on Computer and Information Science (ICIS)*, pages 105–108. IEEE.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Shuo Huang, Zhuang Li, Lizhen Qu, and Lei Pan. 2021. On robustness of neural semantic parsers. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3333–3342.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Xiaonan Li, Hang Yan, Xipeng Qiu, and Xuan-Jing Huang. 2020. Flat: Chinese ner using flat-lattice transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6836–6842.

Zaijing Li, Ting-En Lin, Yuchuan Wu, Meng Liu, Fengxiao Tang, Ming Zhao, and Yongbin Li. 2023. Unisa: Unified generative framework for sentiment analysis. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6132–6142.


(c) 2024 中国中文信息学会计算语言学专业委员会

Bin Liang, Hongcheng Li, Miaoqiang Su, Pan Bian, Xirong Li, and Wenchang Shi. 2018. Deep text classification can be fooled. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4208–4215.

Bill Yuchen Lin, Wenyang Gao, Jun Yan, Ryan Moreno, and Xiang Ren. 2021. Rockner: A simple method to create adversarial examples for evaluating the robustness of named entity recognition models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3728–3737.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ruotian Ma, Xin Zhou, Tao Gui, Yiding Tan, Linyang Li, Qi Zhang, and Xuan-Jing Huang. 2022a. Template-free prompt tuning for few-shot ner. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5721–5732.

Yubo Ma, Zehao Wang, Yixin Cao, Mukai Li, Meiqi Chen, Kun Wang, and Jing Shao. 2022b. Prompt for extraction? paie: Prompting argument interaction for event argument extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6759–6774.

Sewon Min, Kenton Lee, Ming-Wei Chang, Kristina Toutanova, and Hannaneh Hajishirzi. 2021. Joint passage ranking for diverse multi-answer retrieval. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6997–7008.

Marcin Namysl, Sven Behnke, and Joachim Köhler. 2020. Nat: Noise-aware training for robust neural sequence labeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1501–1517.

Nanyun Peng and Mark Dredze. 2015. Named entity recognition for chinese social media with jointly trained embeddings. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 548–554.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using ontonotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152.

Akshay Srinivasan and Sowmya Vajjala. 2023. A multilingual evaluation of ner robustness to adversarial inputs. In *Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023)*, pages 40–53.

Xingwei Tan, Gabriele Pergola, and Yulan He. 2021. Extracting event temporal relations via hyperbolic geometry. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8065–8077. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Kai Wang, Junfeng Tian, Rui Wang, Xiaojun Quan, and Jianxing Yu. 2020. Multi-domain dialogue acts and response co-generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7125–7134.

Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. 2021. Adversarial glue: A multi-task benchmark for robustness evaluation of language models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.


(c) 2024 中国中文信息学会计算语言学专业委员会

Jiayi Wang, Rongzhou Bao, Zhuosheng Zhang, and Hai Zhao. 2022a. Distinguishing non-natural from natural adversarial samples for more robust pre-trained language model. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 905–915.

Xiaolei Wang, Kun Zhou, Ji-Rong Wen, and Wayne Xin Zhao. 2022b. Towards unified conversational recommender systems via knowledge-enhanced prompt learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1929–1937.

Jindong Wang, HU Xixu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Wei Ye, Haojun Huang, Xiubo Geng, et al. 2023. On the robustness of chatgpt: An adversarial and out-of-distribution perspective. In *ICLR 2023 Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models*.

Hongqiu Wu, Ruixue Ding, Hai Zhao, Pengjun Xie, Fei Huang, and Min Zhang. 2023. Adversarial self-attention for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13727–13735.

Hang Yan, Bocao Deng, Xiaonan Li, and Xipeng Qiu. 2019. Tener: adapting transformer encoder for named entity recognition. *arXiv preprint arXiv:1911.04474*.

Yue Zhang and Jie Yang. 2018. Chinese ner using lattice lstm. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1554–1564.

Jiaming Zhang, Jitao Sang, Qi Yi, Yunfan Yang, Huiwen Dong, and Jian Yu. 2023. Imagenet pre-training also transfers non-robustness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3436–3444.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Neil Zhenqiang Gong, Yue Zhang, et al. 2023. Promptbench: Towards evaluating the robustness of large language models on adversarial prompts. *arXiv preprint arXiv:2306.04528*.

## 附录.原始扰动

对受到句子级自然扰动影响的任意一段原始序列$\mathbf{X}^o$，令$l^o$为序列$\mathbf{X}^o$长度，设其对应扰动序列为$\mathbf{X}^r$，令$l^r$为序列$\mathbf{X}^r$长度，同时有$l^o < l^r$，且满足$\forall i,\, j \in C_{\mathbb{N}}\left[1,\, l^r\right],\, i \neq j,\, \langle \boldsymbol{v}_i,\, \boldsymbol{v}_j \rangle = 0$。

$$
\begin{aligned}
\cos\left(\boldsymbol{t}_i^o,\, \boldsymbol{t}_i^r\right) &= \cos\left(\frac{\sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,j}} \boldsymbol{v}_j}{\sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,j}}},\, \frac{\sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,\,j}} \boldsymbol{v}_j + \sum_{j=l^o+1}^{l^r} e^{\hat{\alpha}_{i,\,j}} \boldsymbol{v}_j}{\sum_{j=1}^{l^r} e^{\hat{\alpha}_{i,j}}}\right) \\
&= \cos\left(\sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,j}} \boldsymbol{v}_j,\, \sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,\,j}} \boldsymbol{v}_j + \sum_{j=l^o+1}^{l^r} e^{\hat{\alpha}_{i,\,j}} \boldsymbol{v}_j\right) \\
&= \frac{\left\langle \sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,j}} \boldsymbol{v}_j,\, \sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,\,j}} \boldsymbol{v}_j + \sum_{j=l^o+1}^{l^r} e^{\hat{\alpha}_{i,\,j}} \boldsymbol{v}_j \right\rangle}{\left\| \sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,j}} \boldsymbol{v}_j \right\| \left\| \sum_{j=1}^{l^r} e^{\hat{\alpha}_{i,\,j}} \boldsymbol{v}_j \right\|} \\
&= \frac{\left\langle \sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,j}} \boldsymbol{v}_j,\, \sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,\,j}} \boldsymbol{v}_j \right\rangle + \left\langle \sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,j}} \boldsymbol{v}_j,\, \sum_{j=l^o+1}^{l^r} e^{\hat{\alpha}_{i,\,j}} \boldsymbol{v}_j \right\rangle}{\left\| \sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,j}} \boldsymbol{v}_j \right\| \left\| \sum_{j=1}^{l^r} e^{\hat{\alpha}_{i,\,j}} \boldsymbol{v}_j \right\|} \\
&= \frac{\left\langle \sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,j}} \boldsymbol{v}_j,\, \sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,\,j}} \boldsymbol{v}_j \right\rangle}{\left\| \sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,j}} \boldsymbol{v}_j \right\| \left\| \sum_{j=1}^{l^r} e^{\hat{\alpha}_{i,\,j}} \boldsymbol{v}_j \right\|} = \frac{\left\| \sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,j}} \boldsymbol{v}_j \right\|^2}{\left\| \sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,j}} \boldsymbol{v}_j \right\| \left\| \sum_{j=1}^{l^r} e^{\hat{\alpha}_{i,\,j}} \boldsymbol{v}_j \right\|} \\
&= \frac{\left\| \sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,j}} \boldsymbol{v}_j \right\|}{\left\| \sum_{j=1}^{l^r} e^{\hat{\alpha}_{i,\,j}} \boldsymbol{v}_j \right\|} = \sqrt{1 - \frac{\left\| \sum_{j=l^o+1}^{l^r} e^{\hat{\alpha}_{i,j}} \boldsymbol{v}_j \right\|^2}{\left\| \sum_{j=1}^{l^r} e^{\hat{\alpha}_{i,\,j}} \boldsymbol{v}_j \right\|^2}}
\end{aligned}
$$

若将约束矩阵$\mathbf{W}_{\mathrm{x}}$加入自注意力模块中，设此时过自注意力模块后输出向量为$\tilde{\boldsymbol{t}}_i^o$和$\tilde{\boldsymbol{t}}_i^r$，要证使用约束矩阵$\mathbf{W}_{\mathrm{x}}$可改善模型在句子级自然扰动下的鲁棒性，即证$\cos\left(\tilde{\boldsymbol{t}}_i^o,\, \tilde{\boldsymbol{t}}_i^r\right) > \cos\left(\boldsymbol{t}_i^o,\, \boldsymbol{t}_i^r\right)$。

CCL 2024

## 附录.定理1证明

$$\begin{aligned}
\cos\left(\tilde{\boldsymbol{t}}_i^o,\, \tilde{\boldsymbol{t}}_i^r\right) &= \cos\left(\frac{\sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,j}}\boldsymbol{v}_j + \sum_{j=b_i}^{e_i} e^{\beta\hat{\alpha}_{i,j}}\boldsymbol{v}_j}{\sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,j}} + \sum_{j=b_i}^{e_i} e^{\beta\hat{\alpha}_{i,j}}}, \frac{\sum_{j=1}^{l^r} e^{\hat{\alpha}_{i,\,j}}\boldsymbol{v}_j + \sum_{j=b_i}^{e_i} e^{\beta\hat{\alpha}_{i,j}}\boldsymbol{v}_j}{\sum_{j=1}^{l^r} e^{\hat{\alpha}_{i,j}} + \sum_{j=b_i}^{e_i} e^{\beta\hat{\alpha}_{i,j}}}\right) \\
&= \cos\left(\sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,j}}\boldsymbol{v}_j + \sum_{j=b_i}^{e_i} e^{\beta\hat{\alpha}_{i,j}}\boldsymbol{v}_j, \sum_{j=1}^{l^r} e^{\hat{\alpha}_{i,\,j}}\boldsymbol{v}_j + \sum_{j=b_i}^{e_i} e^{\beta\hat{\alpha}_{i,j}}\boldsymbol{v}_j\right) \\
&= \frac{\left\langle \sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,j}}\boldsymbol{v}_j + \sum_{j=b_i}^{e_i} e^{\beta\hat{\alpha}_{i,j}}\boldsymbol{v}_j, \sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,\,j}}\boldsymbol{v}_j + \sum_{j=l^o+1}^{l^r} e^{\hat{\alpha}_{i,\,j}}\boldsymbol{v}_j + \sum_{j=b_i}^{e_i} e^{\beta\hat{\alpha}_{i,j}}\boldsymbol{v}_j \right\rangle}{\left\| \sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,j}}\boldsymbol{v}_j + \sum_{j=b_i}^{e_i} e^{\beta\hat{\alpha}_{i,j}}\boldsymbol{v}_j \right\| \left\| \sum_{j=1}^{l^r} e^{\hat{\alpha}_{i,\,j}}\boldsymbol{v}_j + \sum_{j=b_i}^{e_i} e^{\beta\hat{\alpha}_{i,j}}\boldsymbol{v}_j \right\|} \\
&= \frac{\left\langle \sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,j}}\boldsymbol{v}_j + \sum_{j=b_i}^{e_i} e^{\beta\hat{\alpha}_{i,j}}\boldsymbol{v}_j, \sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,j}}\boldsymbol{v}_j + \sum_{j=b_i}^{e_i} e^{\beta\hat{\alpha}_{i,j}}\boldsymbol{v}_j \right\rangle}{\left\| \sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,j}}\boldsymbol{v}_j + \sum_{j=b_i}^{e_i} e^{\beta\hat{\alpha}_{i,j}}\boldsymbol{v}_j \right\| \left\| \sum_{j=1}^{l^r} e^{\hat{\alpha}_{i,\,j}}\boldsymbol{v}_j + \sum_{j=b_i}^{e_i} e^{\beta\hat{\alpha}_{i,j}}\boldsymbol{v}_j \right\|} \\
&= \frac{\left\| \sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,j}}\boldsymbol{v}_j + \sum_{j=b_i}^{e_i} e^{\beta\hat{\alpha}_{i,j}}\boldsymbol{v}_j \right\|}{\left\| \sum_{j=1}^{l^r} e^{\hat{\alpha}_{i,\,j}}\boldsymbol{v}_j + \sum_{j=b_i}^{e_i} e^{\beta\hat{\alpha}_{i,j}}\boldsymbol{v}_j \right\|} = \sqrt{1 - \frac{\left\| \sum_{j=l_0+1}^{l^r} e^{\hat{\alpha}_{i,j}}\boldsymbol{v}_j \right\|^2}{\left\| \sum_{j=1}^{l^r} e^{\hat{\alpha}_{i,\,j}}\boldsymbol{v}_j + \sum_{j=b_i}^{e_i} e^{\beta\hat{\alpha}_{i,j}}\boldsymbol{v}_j \right\|^2}}
\end{aligned}$$

由于，$\sqrt{1 - \frac{\left\| \sum_{j=l_0+1}^{l^r} e^{\hat{\alpha}_{i,j}}\boldsymbol{v}_j \right\|^2}{\left\| \sum_{j=1}^{l^r} e^{\hat{\alpha}_{i,\,j}}\boldsymbol{v}_j + \sum_{j=b_i}^{e_i} e^{\beta\hat{\alpha}_{i,j}}\boldsymbol{v}_j \right\|^2}} > \sqrt{1 - \frac{\left\| \sum_{j=l^o+1}^{l^r} e^{\hat{\alpha}_{i,j}}\boldsymbol{v}_j \right\|^2}{\left\| \sum_{j=1}^{l^r} e^{\hat{\alpha}_{i,\,j}}\boldsymbol{v}_j \right\|^2}} = \cos\left(\boldsymbol{t}_i^o,\, \boldsymbol{t}_i^r\right)$。因此，可证：$\cos\left(\tilde{\boldsymbol{t}}_i^o,\, \tilde{\boldsymbol{t}}_i^r\right) > \cos\left(\boldsymbol{t}_i^o,\, \boldsymbol{t}_i^r\right)$。

## 附录.定理2证明

$$\begin{aligned}
\cos\left(\tilde{\boldsymbol{t}}_i^o,\, \tilde{\boldsymbol{t}}_i^r\right) &= \cos\left(\frac{\sum_{j=l^o-w}^{l^o} e^{\hat{\alpha}_{i,j}}\boldsymbol{v}_j}{\sum_{j=l^o-w}^{l^o} e^{\hat{\alpha}_{i,j}}}, \frac{\sum_{j=l^o-w}^{l^o} e^{\hat{\alpha}_{i,\,j}}\boldsymbol{v}_j + \sum_{j=l^o+1}^{l^o+w} e^{\hat{\alpha}_{i,\,j}}\boldsymbol{v}_j}{\sum_{j=l^o-w}^{l^o+w} e^{\hat{\alpha}_{i,\,j}}}\right) \\
&= \cos\left(\sum_{j=l^o-w}^{l^o} e^{\hat{\alpha}_{i,j}}\boldsymbol{v}_j, \sum_{j=l^o-w}^{l^o} e^{\hat{\alpha}_{i,\,j}}\boldsymbol{v}_j + \sum_{j=l^o+1}^{l^o+w} e^{\hat{\alpha}_{i,\,j}}\boldsymbol{v}_j\right) \\
&= \frac{\left\langle \sum_{j=l^o-w}^{l^o} e^{\hat{\alpha}_{i,j}}\boldsymbol{v}_j, \sum_{j=l^o-w}^{l^o} e^{\hat{\alpha}_{i,\,j}}\boldsymbol{v}_j \right\rangle + \left\langle \sum_{j=l^o-w}^{l^o} e^{\hat{\alpha}_{i,j}}\boldsymbol{v}_j, \sum_{j=l^o+1}^{l^o+w} e^{\hat{\alpha}_{i,\,j}}\boldsymbol{v}_j \right\rangle}{\left\| \sum_{j=l^o-w}^{l^o} e^{\hat{\alpha}_{i,j}}\boldsymbol{v}_j \right\| \left\| \sum_{j=l^o-w}^{l^o+w} e^{\hat{\alpha}_{i,\,j}}\boldsymbol{v}_j \right\|} \\
&= \frac{\left\langle \sum_{j=l^o-w}^{l^o} e^{\hat{\alpha}_{i,j}}\boldsymbol{v}_j, \sum_{j=l^o-w}^{l^o} e^{\hat{\alpha}_{i,\,j}}\boldsymbol{v}_j \right\rangle}{\left\| \sum_{j=l^o-w}^{l^o} e^{\hat{\alpha}_{i,j}}\boldsymbol{v}_j \right\| \left\| \sum_{j=l^o-w}^{l^o+w} e^{\hat{\alpha}_{i,\,j}}\boldsymbol{v}_j \right\|} \\
&= \frac{\left\| \sum_{j=l^o-w}^{l^o} e^{\hat{\alpha}_{i,j}}\boldsymbol{v}_j \right\|^2}{\left\| \sum_{j=l^o-w}^{l^o} e^{\hat{\alpha}_{i,j}}\boldsymbol{v}_j \right\| \left\| \sum_{j=l^o-w}^{l^o+w} e^{\hat{\alpha}_{i,\,j}}\boldsymbol{v}_j \right\|} \\
&= \frac{\left\| \sum_{j=l^o-w}^{l^o} e^{\hat{\alpha}_{i,j}}\boldsymbol{v}_j \right\|}{\left\| \sum_{j=l^o-w}^{l^o+w} e^{\hat{\alpha}_{i,\,j}}\boldsymbol{v}_j \right\|} = \sqrt{1 - \frac{\left\| \sum_{j=l^o+1}^{l^o+w} e^{\hat{\alpha}_{i,j}}\boldsymbol{v}_j \right\|^2}{\left\| \sum_{j=l^o-w}^{l^o+w} e^{\hat{\alpha}_{i,\,j}}\boldsymbol{v}_j \right\|^2}}
\end{aligned}$$

由于，$\lim_{w\to 0}\sqrt{1 - \frac{\left\| \sum_{j=l^o+1}^{l^o+w} e^{\hat{\alpha}_{i,j}}\boldsymbol{v}_j \right\|^2}{\left\| \sum_{j=l^o-w}^{l^o+w} e^{\hat{\alpha}_{i,\,j}}\boldsymbol{v}_j \right\|^2}} > \sqrt{1 - \frac{\left\| \sum_{j=l^o+1}^{l^r} e^{\hat{\alpha}_{i,j}}\boldsymbol{v}_j \right\|^2}{\left\| \sum_{j=1}^{l^r} e^{\hat{\alpha}_{i,\,j}}\boldsymbol{v}_j \right\|^2}} = \cos\left(\boldsymbol{t}_i^o,\, \boldsymbol{t}_i^r\right)$。因此，可证：$\cos\left(\tilde{\boldsymbol{t}}_i^o,\, \tilde{\boldsymbol{t}}_i^r\right) > \cos\left(\boldsymbol{t}_i^o,\, \boldsymbol{t}_i^r\right)$。

CCL 2024

# 银瞳：基于自适应语义空间学习的中文金融多任务大模型

**周宇航[1], 李泽平[1], 田思雨[1], 倪雨琛[2], 张健[3], 刘响[4], 叶广楠[1], 吴杰[1], 柴洪峰[1]**
[1]复旦大学 计算机科学技术学院 金融科技研究院，上海
[2]同济大学电子与信息工程学院，上海
[3]达观数据，上海
[4]纽约大学坦登工学院，纽约
{yuhangzhou22, zepingli23, sytian23}@m.fudan.edu.cn
2230733@tongji.edu.cn; zhangjian@datagrand.com; xl493@nyu.edu
{yegn, jwu, hfchai}@fudan.edu.cn

## 摘要

大语言模型正逐渐被用于各种垂直领域，利用其广泛的知识储备来赋能领域中的多种场景。然而，各领域拥有多种待学习的特定任务，且多源异构的领域数据容易引发模型进行任务迁移时的冲突。基于此，本研究提出自适应语义空间学习框架，利用对语义空间内数据的自适应重分布，提升多专家模型的性能及选择效果，并基于此框架训练了一个金融多任务大模型“银瞳”。研究结果表明，我们的框架只需利用10%的数据就能达到接近全数据训练的效果，并拥有较强的泛化表现。

**关键词：** 金融大模型；自适应语义空间学习；多任务学习

# SilverSight: A Multi-Task Chinese Financial Large Language Model Based on Adaptive Semantic Space Learning

**Yuhang Zhou[1,2] Zeping Li[1,2] Siyu Tian[1,2] Yuchen Ni[3]**
**Jian Zhang[4] Xiang Liu[5] Guangnan Ye[1,2] Jie Wu[1,2] Hongfeng Chai[1,2]**

[1]Institute of FinTech, Fudan University
[2]School of Computer Science, Fudan University
[3]School of Electronics and Information Engineering, Tongji University
[4]DataGrand Inc.
[5]Tandon School of Engineering, New York University

## Abstract

Large language models (LLMs) are increasingly being applied across various specialized fields, leveraging their extensive knowledge to empower a multitude of scenarios within these domains. However, each field encompasses a variety of specific tasks that require learning, and the diverse, heterogeneous data across these domains can lead to conflicts during model task transfer. In response to this challenge, our study introduces an Adaptive Semantic Space Learning (ASSL) framework, which utilizes the adaptive reorganization of data distributions within the semantic space to enhance the performance and selection efficacy of multi-expert models. Utilizing this framework, we trained a financial multi-task LLM named ”SilverSight”. Our research findings demonstrate that our framework can achieve results close to those obtained with full data training using only 10% of the data, while also exhibiting strong generalization capabilities.

**Keywords:** Financial LLMs , Adaptive Semantic Space Learning , Multi-task Learning

©2024 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版

第二十三届中国计算语言学大会论文集，第955页-第972页，太原，中国，2024年7月25日至28日。
卷1：主会论文
(c) 2024 中国中文信息学会计算语言学专业委员会

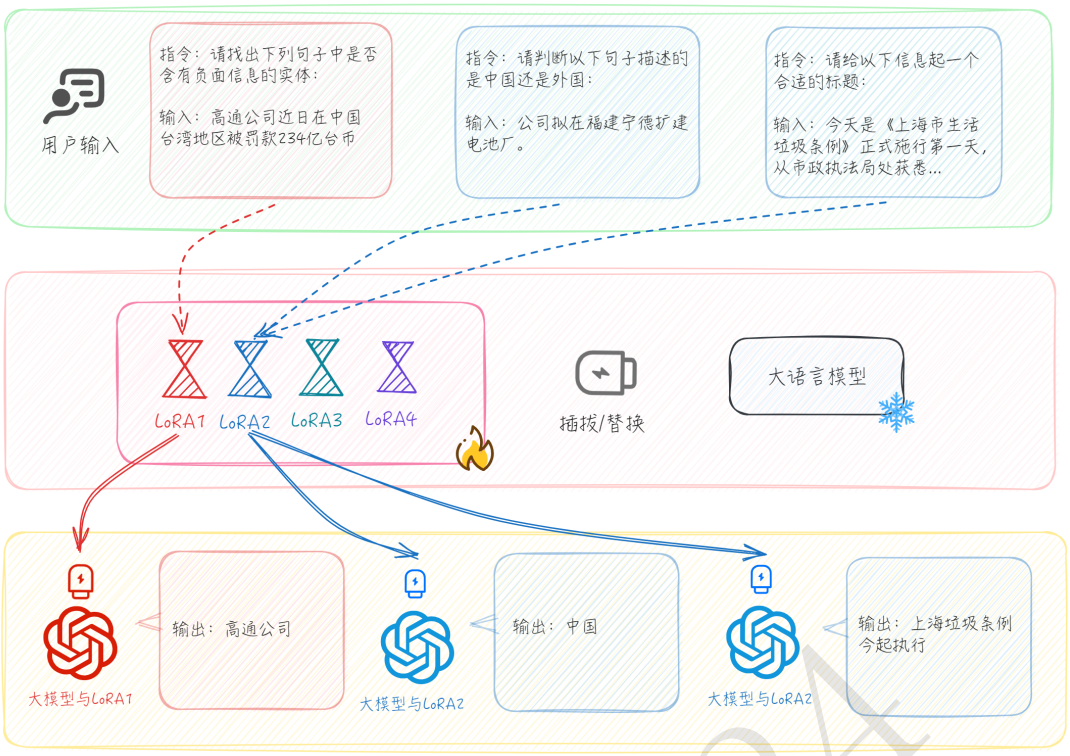



Figure 1: “银瞳”金融多任务大模型流程示意

## 1 引言

近期，大型语言模型如GPT-4 (OpenAI, 2023)和LLaMA (Touvron et al., 2023)等，在自然语言处理领域各项任务中展现了惊人的实力。这些模型凭借强大的知识存储和上下文理解能力，其优异性能在金融 (Zhou et al., 2024; Chen et al., 2023)、法律 (Cui et al., 2023)等多个专业领域已得到广泛应用。在构建大语言模型的过程中，指令微调技术扮演了核心角色，它利用有监督的数据集对模型进行微调，助力模型从简单的文本续写向复杂任务解答进行风格跃迁 (Qiu et al., 2020)。然而，受制于特定领域的数据量和计算资源限制，指令微调往往需要借助参数高效的微调策略 (Houlsby et al., 2019)，这些策略仅需对基础模型进行少量参数的更新或增加，便能显著提升模型对指令的响应能力。在此背景下，低秩自适应(Low-Rank Adaptation, LoRA) 方法 (Hu et al., 2021)应运而生，通过为原始模型引入可分解的低参数量旁路矩阵，实现了模块化改造及性能的全面提升，并具备了简便的“即插即用”特性。

当前，自然语言处理领域的研究热点之一是采用混合专家模型(Mixture of Experts, MoE) (Jordan and Jacobs, 1994)来解决多任务处理的挑战。混合专家模型是一种集成学习策略，它通过整合多个专家模型，利用每个模型在子问题上的专长，在处理复杂任务时提升整体性能 (Dou et al., 2023)。Zadouri (Zadouri et al., 2023)首次提出将LoRA方法与MoE框架相结合，采用基于token级别的软路由机制，对各个专家模型的输出进行加权融合，有效提升了模型的性能。近期，研究者进一步提出了LoRAMoE方法 (Dou et al., 2023)，该方法将专家模型划分为两组：一组专注于通用世界知识的处理，另一组则致力于在学习指令微调过程中遇到的新任务。这种设计旨在增强模型处理下游任务的能力，同时保留大模型积累的世界知识。这些多专家策略在处理不同问题时展现出了优势，但它们对于token级别的处理以及训练后专家和路由策略的固定性，限制了模型快速适应新指令或场景的能力。为了解决这一问题，研究人员提出了LoraRetriever方法 (Zhao et al., 2024)，该方法借鉴了检索思想，将每个专家模型中的12条数据嵌入的平均值作为LoRA的自身嵌入，并利用输入问题的嵌入来检索最匹配的LoRA，从而有效弥补了多专家模型在动态增删方面的不足，进一步提升了混合训练模型的性能和泛化能力。

尽管上述方法在提升LoRA等适配器的动态适应性方面表现出色，但它们在训练LoRA时往往依赖于任务类型对训练数据进行手动划分，并未充分考虑数据在语义空间中的内在联系。这


(c) 2024 中国中文信息学会计算语言学专业委员会

种处理方式可能导致LoRA专家的嵌入产生偏差，从而影响系统的整体性能。与此同时，来自不同来源的数据在多样性和质量上存在差异 (Touvron et al., 2023)，而如何合理配置不同数据量的比例在先前的研究中并未得到充分探讨。这进一步加剧了LoRA专家选择和数据选择的问题，使其成为一个亟待解决的关键挑战。因此，探索一种能够同时优化LoRA专家选择和数据选择策略的方法，以保持数据在语义空间中的连续性和关联性，成为了提升混合专家模型性能的重要研究方向。这要求我们不仅要关注LoRA的训练过程，还要重视数据的分布修正机制，以确保模型能够在面对复杂多变的任务时，依然能够保持高效和准确的性能表现。

为了填补现有研究的空白，我们提出了自适应语义空间学习框架(Adaptive Semantic Space Learning, ASSL)，并通过金融多任务数据集训练了一个名为“银瞳” (SilverSight) 的金融多任务大语言模型。ASSL框架利用多任务训练数据在语义空间中的相似性进行聚类，从而摒弃了对训练数据任务类型的预定义。实验结果表明，与基于预设任务类型的分类方法相比，我们的方法具有较为显著的优势，确保了每个专家模型能够被分配到与其最相关的下游任务中。同时，我们发现不同来源和格式的训练任务可能引发不可避免的任务冲突，影响模型的指令遵循能力。此外，通过利用语义空间的相似性，我们的方法能够聚合互补任务的多种数据，从而在相关任务上提升模型性能。在ASSL框架下，我们在每个聚类中采用基于模型自我进化的数据重分布策略，进行自适应的数据挑选。通过两阶段的数据自适应筛选，我们确保了用于训练每个LoRA专家的少量数据具备高质量、高覆盖率以及高必要性。在原始数据质量及数量分布不均的情况下，ASSL框架通过考虑语义空间分布密度和模型自反馈机制，对每个LoRA的训练数据进行自适应筛选。这种方法使得模型能够对长尾分布中的头部相似性数据和尾部小样本数据进行更均衡的拟合，从而提高了模型在多样任务上的泛化能力和性能表现。

我们从金融领域的23个不同来源收集了22万条中文金融微调数据，并将这些数据分类为情感分析、金融问答、文本生成、金融选择题等多个任务类型。基于这些实验数据，我们得出以下主要发现：（1）通过利用语义空间中的相似性进行聚类，我们能够识别出相互促进和存在冲突的训练任务。采用多个专家模型分别针对特定领域任务进行学习，可以使得每个专家专注于其擅长的领域，实现“各司其职”的效果。（2）结合数据在语义空间中的密度分布和模型自身对训练数据的需求，我们能够有效地对数据进行语义平滑重分布。这种方法使得整个系统能够在仅使用10%的数据进行微调的情况下，达到与使用全量数据进行微调相似的效果。（3）通过对聚类内部的数据进行分布平滑处理，我们利用聚类内数据嵌入的质心作为LoRA专家的嵌入，从而实现了LoRA选择的最优化。

这些发现不仅验证了ASSL框架在金融多任务学习中的有效性，而且为如何高效利用有限数据资源、提高模型性能提供了新的视角和方法。这些成果对于推动领域自然语言处理技术的发展具有重要意义。我们的主要贡献如下：

1. 我们创新性地提出了自适应语义空间学习(ASSL) 框架，通过分析数据在语义空间的分布特性，实现了多任务场景下LoRA专家与数据的自适应选择机制。此举有效调整了专家与训练数据的分布平衡，借助语义空间信息对训练数据进行平滑处理，从而避免了专家过拟合或训练过程中的冲突现象。

2. 我们在金融领域对ASSL框架的有效性进行了实证验证，并成功训练了一个中文金融多任务大语言模型“银瞳”。在两个评测基准上，我们对模型的泛化能力及多任务表现进行了全面测试，充分展现了该框架的优越性和应用潜力。

3. 我们对模型不同能力的来源进行了深入的多维度分析，为未来专家与数据自适应选择领域的研究提供了新的思考方向。

## 2 准备工作

### 2.1 LoRA

对大语言模型进行全参数微调需要耗费大量的数据和计算资源，并且可能会导致灾难性遗忘的问题 (Luo et al., 2023)。为了解决这一问题，近期研究人员开发了多种高效的微调方法，其中LoRA (Hu et al., 2021)因其有效性而被广泛应用。LoRA方法利用低秩矩阵分解的思想来减少微调过程中需要调节的参数量，通过在网络中引入一个附加的旁路结构$W_0^{m\times n}$ 来实现。具


(c) 2024 中国中文信息学会计算语言学专业委员会

体来说，LoRA通过增加$W_0^{m\times n}$ 到原始权重矩阵$W^{m\times n}$ 上，来修改网络中的权重，即：

$$W'^{m\times n} = W^{m\times n} + W_0^{m\times n} = W^{m\times n} + A^{m\times r}B^{r\times n} \tag{1}$$

其中，$A^{m\times r}$ 和$B^{r\times n}$ 是从$W_0^{m\times n}$ 通过低秩分解得到的两个矩阵，r为矩阵的秩。当输入模型的是token表征$x^{n\times 1}$ 时，模型的输出$y^{m\times 1}$ 可以表示为：

$$y = W'x = (W + AB)\,x = Wx + ABx \tag{2}$$

这里，$x^{n\times 1}$ 代表模型输入的token表征，而$y^{m\times 1}$ 代表模型输出。

### 2.2 任务定义

在ASSL框架中，我们利用数据在语义空间的分布情况，设计了一个多专家系统。主要关注两个核心任务：专家自适应选择和数据自适应选择。这两个任务旨在提高系统在处理多任务环境下的灵活性和效率。

专家自适应选择任务侧重于在给定的输入情况下，如何自适应地挑选最合适的LoRA专家来处理特定的任务。考虑到本框架的多任务系统是一个由$n$种不同任务组成的集合$T = \{t_1, t_2, \ldots, t_n\}$，以及$m$个不同的LoRA专家构成的集合$E = \{e_1, e_2, \ldots, e_m\}$。其中每个专家$e_i$都经过若干特定任务的训练，以优化其性能。对于任意输入$x$，我们的目标是找到一个映射函数$f : X \to E$，使得$f(x)$能够选择出最适合当前输入$x$的专家$e$。这可以通过最大化输入$x$与专家$e$之间相关度的方式来实现，即：

$$e^* = f(x) = \arg\max_{e\in E} \text{Relevance}(x, e) \tag{3}$$

其中，函数Relevance$(x, e)$衡量了输入$x$与专家$e$之间的相关性。

另一方面，数据自适应选择任务专注于处理来自多个源的异构数据，并解决数据中存在的长尾分布问题。我们有一个数据集$D = \{d_1, d_2, \ldots, d_l\}$，每个$d_i$代表不同源的数据。在实际应用中，这些数据往往遵循长尾分布，即大量数据集中在少数类别中，而大部分类别仅包含少量数据。我们的目标是通过一个转换函数$g : D \to D'$来处理原始数据集$D$，将其转换为一个更平滑、分布更均匀的数据集$D'$，从而减轻长尾分布的影响。

## 3 自适应语义空间学习框架

我们提出了一种名为ASSL (自适应语义空间学习) 的框架，该框架旨在实现对LoRA专家及其数据自适应选择的功能。框架示意如 Figure 2 所示，它确保了每个专家能够发挥最佳性能，并使整个系统在多任务集合上拥有卓越的表现。在本章中，我们将详细阐述框架的两个核心组成部分，即如何通过数据在语义空间的重新分布来统一实现专家以及数据之间的自适应选择。

### 3.1 LoRA专家的自适应选择

在处理LoRA专家的自适应选择时，我们的目标是优化数据划分以避免任务间的潜在冲突，并确保对于每个输入，都能匹配到最擅长处理该类问题的专家。这一目标主要分为两个方面：（1）如何更有效地训练多任务指令；（2）如何根据用户输入选择最合适的LoRA专家。

首先，为了增强系统对多样化指令的泛化能力，我们对任务集中的每个子任务进行了指令扩充。具体而言，我们为每个任务人工编写了30条具有相似语义但表述不同的指令。然后，使用句子编码器Emb(·)对每条指令及其输入数据的拼接Ins $\oplus$ Inp 进行编码，获得所有数据在同一语义空间中的嵌入向量。我们采用K-Means聚类方法 (MacQueen and others, 1967)将语义空间中相近的句子聚集成$K$ 类，以此来优化任务的数据划分。聚类过程可以通过以下公式表示：

$$\text{Cluster}_k = \arg\min_{x\in X} \sum_{i=1}^{K} \sum_{x\in C_i} \|x - \mu_i\|^2 \tag{4}$$

其中，$X$ 表示所有数据点的集合，$C_i$ 是第$i$ 个聚类中的数据点集合，而$\mu_i$ 是第$i$ 个聚类的质心。


(c) 2024 中国中文信息学会计算语言学专业委员会

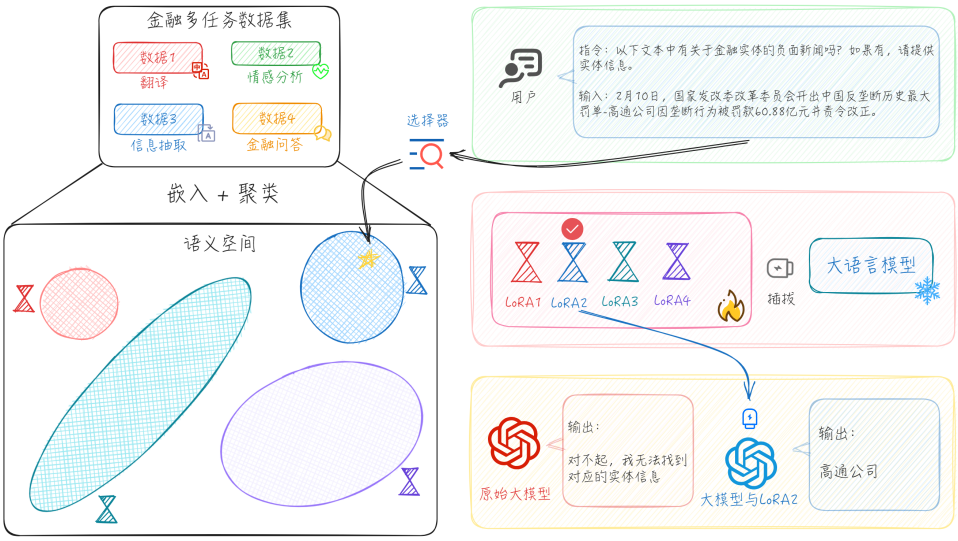



Figure 2: 自适应语义空间学习框架

通过 4.5 实验证明，与混合任务训练以及按预定义标签进行划分训练的方式相比，这种基于语义的聚类训练法能够显著提升系统性能。针对每个LoRA专家，我们选择其聚类的质心作为该专家的语义嵌入，每个聚类的质心$\mu_i$ 是聚类中所有点语义嵌入的平均位置，计算公式如下：

$$\text{Emb}(e_i) = \mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x \tag{5}$$

其中，$C_i$ 是第$i$ 个聚类中的数据点集合，而$|C_i|$ 表示集合$C_i$ 中元素的数量。$\mu_i$ 代表聚类$i$中所有点的均值，即该聚类的质心。每当有用户输入时，系统通过以下公式寻找与用户输入语义嵌入最接近的专家进行响应：

$$e^* = \arg\min_{e \in E} \|\text{Emb}(x) - \text{Emb}(e)\| \tag{6}$$

这里，$\text{Emb}(x)$ 是用户输入的嵌入向量，$\text{Emb}(e)$ 是专家的语义嵌入向量，而$e^*$ 则是被选中的专家。通过此种匹配方式，系统能够在语义空间中找到与训练任务最匹配的专家。

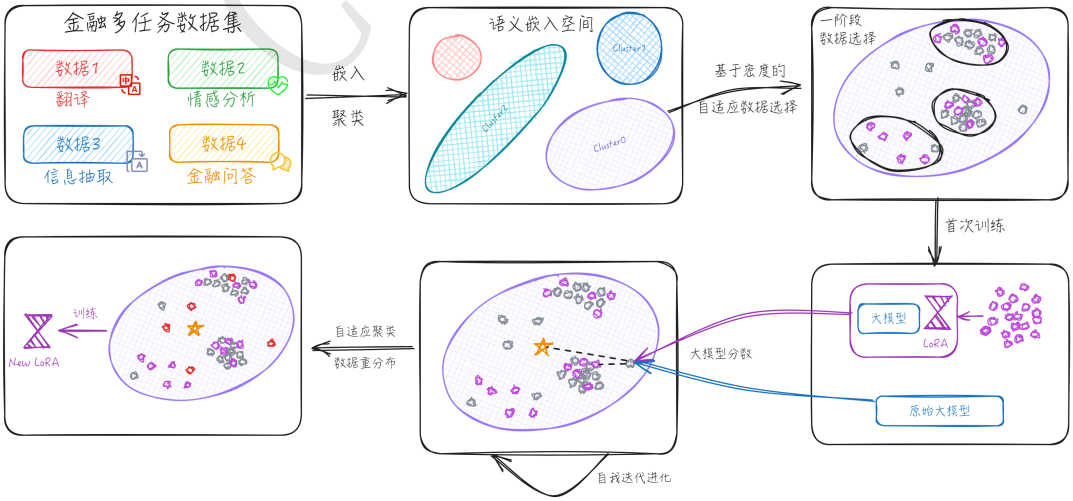



Figure 3: 自适应语义空间数据重分布流程


(c) 2024 中国中文信息学会计算语言学专业委员会

### 3.2 多源异构数据的自适应选择

在本章中，我们讨论如何对多任务有监督数据进行聚类，以隔离相互冲突的任务并聚集相互增强的任务。虽然这种方法有效地解决了任务间的冲突，但它也引入了新的挑战，如数据配比失衡和质量不一，特别是长尾数据分布的问题。为了解决这些问题，我们为每个聚类设计了一个两阶段的数据重分布操作，目的是在保持少量数据集上进行有效微调的同时，达到比拟全数据集微调的效果，具体流程见 Figure 3。

在第一阶段，针对聚类簇中数据失衡的问题，我们基于DBSCAN算法 (Schubert et al., 2017)设计了一种自适应调整的A-DBSCAN算法，对每个聚类中的数据进行嵌套聚类，具体算法流程见 A.1 。该算法可以根据数据密度在不同区域中动态调整连通数量。具体实现步骤如下：首先，算法通过K-最近邻算法(K-Nearest Neighbor, KNN) (Peterson, 2009)距离计算框架来评估数据点在语义空间中的局部密度。每个数据点的局部密度为其至最近的$k$个邻居的平均距离的倒数，其数学表达式为：

$$\rho_i = \frac{1}{\frac{1}{k}\sum_{j=1}^{k} d(x_i, x_{ij})} \tag{7}$$

其中，$d(x_i, x_{ij})$表示数据点$x_i$与其第$j$个最近邻居$x_{ij}$之间的距离。

随后，算法根据计算得到的局部密度值将数据点进行排序，形成一个优先队列，优先处理局部密度较高的数据点。在每次迭代中，选取队列中局部密度最高的点作为起始点，围绕此点进行簇的生成。

在自适应过程中，算法利用队列中所有数据点KNN距离的中位数定义邻域半径$\varepsilon$。同时，启发式地将邻域节点数$MinPts$的初始值设置为：

$$MinPts_{\text{init}} = \frac{\varepsilon \times \rho_{\max}}{2} \tag{8}$$

其中，$\rho_{\max}$表示全局最大局部密度，即初始优先队列中第一个数据点的局部密度。在每个簇形成后，$MinPts$根据以下公式进行更新，以适应当前的局部密度环境：

$$MinPts_{\text{update}} = \max\left(2, \frac{\rho_{\text{current}}}{\rho_{\max}} \times MinPts\right) \tag{9}$$

此处，$\rho_{\text{current}}$表示当前优先队列中第一个点的局部密度。这种动态调整策略使得算法能够更灵活地适应不同密度的数据分布，提高聚类的准确性和效率。同时，这一策略允许我们在每个子簇中平均选取数据量，同时过滤掉不连通的噪声点，有效地对高密度数据进行下采样，对低密度数据进行上采样，从而避免过拟合和欠拟合。

在第二阶段，我们利用第一阶段筛选出的少量数据对模型进行初步的LoRA微调。参考先前的研究，大语言模型在微调阶段主要学习新的语言风格分布，而难以获得新的领域知识 (Ren et al., 2024)。因此，我们考虑数据与大模型自身知识的冲突情况，根据初步训练前后大模型对未选数据的得分差异，设计了两种得分机制来评估每条数据对当前模型的价值。差异得分和比例得分分别定义为：

$$\text{Score}_{\text{diff}}(x) = \text{LLM}_{\text{Raw}}(x) - \text{LLM}_{\text{LoRA}}(x) \tag{10}$$

$$\text{Score}_{\text{prop}}(x) = -\frac{\text{LLM}_{\text{LoRA}}(x)}{\text{LLM}_{\text{Raw}}(x) + 1} \tag{11}$$

$$\text{Score}_{\text{llm}}(x) = \text{Score}_{\text{diff}}(x) + \text{Score}_{\text{prop}}(x) \tag{12}$$

其中，$\text{Score}_{\text{Raw}}(x)$以及$\text{Score}_{\text{LoRA}}(x)$通过Rouge方法 (Lin, 2004)计算。此外，为了确保新选数据的质量和对整个聚类簇的覆盖度，受启发于MMR公式 (Carbonell and Goldstein, 1998)，我们设计了训练数据的效用函数：

$$\text{U}(d_{\text{new}}) = \lambda_1 \cdot \text{sim}(\mu, d_{\text{new}}) - \lambda_2 \cdot \max_{d \in D_{\text{selected}}} \text{sim}(d, d_{\text{new}}) + \lambda_3 \cdot \text{Score}_{\text{llm}}(d_{\text{new}}) \tag{13}$$


(c) 2024 中国中文信息学会计算语言学专业委员会

其中，$\mu$ 代表聚类中心点，$d_{\text{new}}$ 代表待加入的数据点，$D_{\text{selected}}$ 表示已经被选中的数据点集合，$\lambda_1$，$\lambda_2$ 和$\lambda_3$ 是三个可设定的权重参数，用于调整相似度、多样性和模型得分对最终数据点效用值的贡献。

经过两阶段的数据筛选，每个聚类中的数据将在语义空间中进行重分布，从而促使模型学习那些罕见但有益的数据，避免在常见数据集上产生过拟合现象。

## 4 实验

本章中，我们将利用金融领域的公开数据集，训练一个中文金融多任务大模型“银瞳”，利用两个中文金融的评测分析来验证本研究提出的自适应语义空间学习算法的有效性。

### 4.1 数据介绍

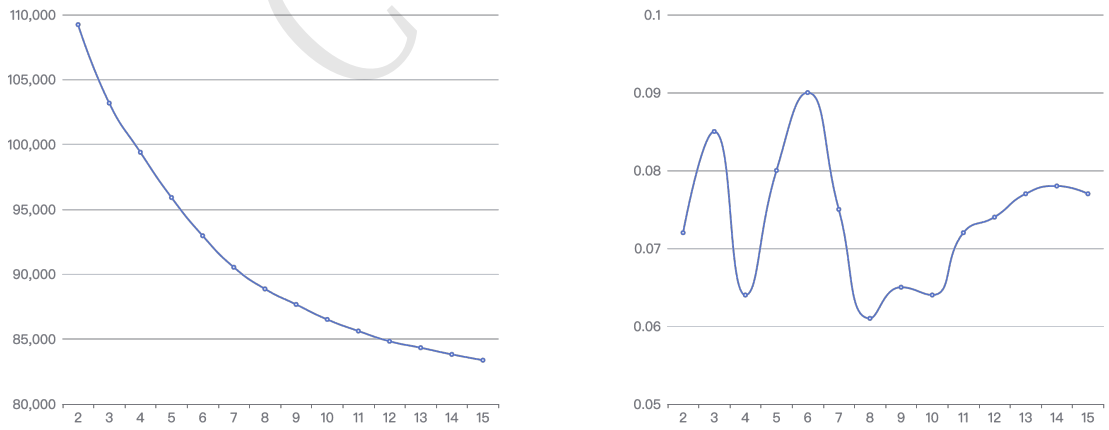
我们从金融领域23个不同的来源收集了约22万条中文金融微调数据，并将这些数据分类为情感分析、金融问答、金融问答等7类任务，具体介绍见 A.2。同时，我们使用CFLEB (Lu et al., 2023)和FinEval (Zhang et al., 2023)评测集作为评估工具，旨在评估大语言模型在金融领域的知识储备、指令跟随及任务执行能力。

CFLEB评测集是利用公开的研报及新闻等项目构建的高质量、高实用性的评测基准，包含了六种自然语言处理任务，衡量金融领域模型在情感分析、问答、摘要生成、信息抽取、语义匹配等方面理解与生成的全面能力。

FinEval评测集是针对金融领域知识评估而设计的一套全面数据集，包含了共4661道选择题。这些题目覆盖了34个不同的学术领域，如金融学、保险学、宏观经济学和税法等，主要分为四大类别：金融、经济、会计以及资格证考试，以全面测试大型模型对金融领域通用知识的掌握程度，评估其在金融专业领域内的先进知识与实际应用能力。

### 4.2 实验设置

我们使用了具有代表性的中文主导模型Qwen1.5-7B-Chat (Bai et al., 2023)作为基座模型，在2张A800显卡上进行实验。对于每一个专家，都为其训练一个LoRA适配器，其中r设置为16，$\alpha$设计为32。对于所有的微调，我们设置学习率为1e-4，经过10% 步的预热，再利用余弦衰减方式将其减小至1e-5，总共经过3轮的训练。在K-Means聚类算法中，经过我们的多次尝试，并利用误差平方和（Sum of Squared Errors, SSE）和轮廓系数（Silhouette Coefficient, SC）对K=2至K=15的所有数据进行测量，如Figure 4所示。最终设置K=6。同时在自适应的A-DBSCAN算法中，我们选取了k=20作为最初的KNN计算参数。对于修正的MMR公式，我们将$\lambda_1$，$\lambda_2$ 和$\lambda_3$ 分别设定为0.2, 0.2以及0.6，以平衡每个数据与聚类中心相似度、多样性以及模型得分对最终效用函数的贡献。




(a) K=2至K=15的误差平方和示意图　　(b) K=2至K=15的轮廓系数示意图

Figure 4: K-Means算法参数K数值挑选图


(c) 2024 中国中文信息学会计算语言学专业委员会

### 4.3 “银瞳”金融多任务大模型

在本研究中，我们利用设计的ASSL框架对数据进行处理，并训练出了针对中文金融领域的多任务大语言模型“银瞳”。首先我们采用了K-Means算法对公开金融领域的数据进行了语义空间聚类，最终形成六个类别，多种类别个数结果的数据分布见 A.4。接下来，我们对每个聚类中的数据进行平滑分布操作：①第一阶段根据自适应的密度聚类算法A-DBSCAN进行数据筛选，在每个聚类中对高密度区域和低密度区域分别进行下采样和上采样，每个聚类选取近2000条微调数据，将处于极低密度区域的离群数据作为噪声。②第二阶段的重分布筛选旨在通过训练前的原始模型以及第一阶段数据训练后的模型，自适应补充在第一阶段中未被选中的必要数据点，使得微调数据分布更加平滑，并增强数据选择的多样性。在每个聚类中总共筛选出4000条数据，合计约占总数据量的10%。我们利用这六类经过自适应语义分布平滑后的数据，分别训练了六个不同的LoRA专家模型，以适应不同的金融领域任务。当处理特定的金融问题时，通过计算问题在同一个语义空间中的表征与六个LoRA专家表征的相似度，自动选择最适合的LoRA专家进行回答，实现了一个高效、自适应且分工明确的金融多任务大模型系统。

Table 1: CFLEB的评估结果：All-data表示使用我们收集的所有数据训练的单个模型，Cluster-data表示使用所有聚类数据训练的多专家系统，SilverSight-mix表示使用所有过滤数据训练的单一模型。在评估结果中，我们使用10%数据训练的SilverSight模型获得了与All-data相似的结果，而Cluster-data系统获得了最佳性能。

| **Model** | **FE** (ACC) | **QA** (F1) | **NA** (Rouge) | **RE** (F1) | **NSP1** (ACC) | **NSP2** (F1) | **NL** (F1) | **AVG** |
|---|---|---|---|---|---|---|---|---|
| Qwen-1.5-7B-Chat | 65.3 | 31.6 | 30.9 | 10.4 | 90.8 | 9.1 | 33.4 | 38.8 |
| All-data | **72.3** | **87.1** | **53.8** | **35.8** | 93.2 | 31.1 | **90.1** | 66.2 |
| Cluster-data | 70.9 | 83.4 | 53.1 | 35.3 | **93.5** | **56.2** | **90.1** | **68.9** |
| PIXIU | 44.7 | 25.4 | 16.1 | 6.5 | 0 | 26.5 | **26.1** | 20.8 |
| FinGPT | 50.6 | 14.8 | 25.9 | 4.3 | 38.6 | 4.7 | 7.1 | 20.9 |
| DISC-FinLLM | **69.3** | **42.4** | **31.0** | **10.1** | **84.6** | **57.9** | 23.4 | **45.5** |
| $SilverSight_{LoRA-0}$ | 68.2 | 60.8 | 32.5 | 10.9 | 83.8 | 10.6 | 45.7 | 44.6 |
| $SilverSight_{LoRA-1}$ | **70.2** | **74.5** | 49.5 | **31.4** | 87.8 | **64.8** | 62.9 | **63.0** |
| $SilverSight_{LoRA-2}$ | 65.4 | 61.9 | 45.2 | 11.9 | 11.8 | 45.4 | 25.9 | 38.2 |
| $SilverSight_{LoRA-3}$ | 67.3 | 53 | 37.7 | 12.4 | **92** | 9.1 | 43.8 | 45.0 |
| $SilverSight_{LoRA-4}$ | 57.5 | 72.9 | **49.6** | 14.9 | 89.2 | 18 | **80.5** | 54.7 |
| $SilverSight_{LoRA-5}$ | 66.6 | 61.8 | 38.4 | 14 | 73.2 | 18.5 | 49.3 | 46.0 |
| SilverSight(our, 10%) | **66.6** | 73.3 | **49.2** | **34.3** | **93.0** | **57.1** | **82.4** | **65.1** |
| SilverSight-mix | 65.7 | **75.4** | 48.3 | 30.4 | 89.2 | 31.1 | 82.3 | 60.3 |

### 4.4 主要结果

Table 1与 Table 2展示了我们对“银瞳”大型模型进行的评测实验结果。在针对CFLEB数据集的评测中，我们注意到FinFE、FinQA、FinNA、FinRE的指标与FinNSP1、FinNSP2的指标之间存在一定的矛盾，往往表现出此消彼长的趋势。然而，我们采用的多专家方法有效地缓解了这种偏差，几乎所有指标都超越了采用混合数据训练的SilverSight-mix模型。此外，我们的方法仅使用总量10%的数据，就能达到与全数据微调模型相媲美的测试成绩，在平均分数上只有1%的差距，在FinNSP2任务上的F1指标甚至为全数据微调模型的2倍。

同时，我们发现对每个聚类进行全数据微调的模型，在CFLEB数据集上的表现相较于全数据微调模型有所提升，这充分证明了ASSL框架的有效性。为了检验模型的泛化能力以及对金融领域知识的掌握，我们在未经过训练的FinEval评测基准上进行了测试。结果显示，训练后的模型在数学计算和资格证考试的选择题上的正确率下降了不到2%，但在经济和金融领域的选择题上的正确率提升了4%，FinEval的平均分数也相较于原始模型有所提高。

为验证该LoRA自适应挑选算法的有效性，我们对聚类后的六个LoRA专家逐一在CFLEB数据集和FinEval数据集上进行测试，如 Table 1与 Table 2所示。实验结果显


(c) 2024 中国中文信息学会计算语言学专业委员会

示，LoRA自适应挑选算法在每个任务上的表现和在该任务上表现最出色的单个LoRA专家相近，证明大多数时候这种LoRA自适应挑选算法均能挑选到最合适的LoRA专家回答问题。同时， Figure 5不仅证明了LoRA自适应挑选算法的有效性，也证明了同时使用聚类方法和LoRA自适应挑选算法能够保证语义空间的一致性与连贯性，最终提升系统的综合性能。

Table 2: FinEval的评估结果

| FinEval | Accounting | Certificate | Economy | Finance | AVG |
|---|---|---|---|---|---|
| Qwen-7B-Chat | 44.5 | 53.6 | 52.1 | 51.5 | 50.5 |
| ChatGPT | 45.2 | 55.1 | 61.6 | 59.3 | 55.0 |
| Qwen-1.5-7B-Chat | **69.5** | **71.3** | 62.8 | 65.6 | 67.8 |
| GPT-4 | 59.3 | 70.4 | **74.5** | **71.0** | **68.6** |
| PIXIU | 36.1 | 34.7 | 37.2 | 34.1 | 35.5 |
| FinGPT | **50.8** | **52.1** | 42.5 | 50.5 | **49.6** |
| DISC-FinLLM | 42.0 | 49.1 | **47.9** | **63.8** | 48.7 |
| $SilverSight_{LoRA-0}$ | **68.9** | **72.2** | 64.7 | 65.9 | **68.3** |
| $SilverSight_{LoRA-1}$ | 66.9 | 70.7 | 59.4 | 62 | 65.3 |
| $SilverSight_{LoRA-2}$ | 62 | 65.6 | 58.5 | 62.6 | 62.6 |
| $SilverSight_{LoRA-3}$ | 67.5 | 69.8 | **67.1** | 67.2 | 68 |
| $SilverSight_{LoRA-4}$ | 68.5 | 70 | 63.8 | 66.2 | 67.5 |
| $SilverSight_{LoRA-5}$ | 68.2 | 71.3 | 63.3 | 65.6 | 67.5 |
| SilverSight(our, 10%) | 67.9 | 70 | 66.7 | **67.9** | **68.3** |

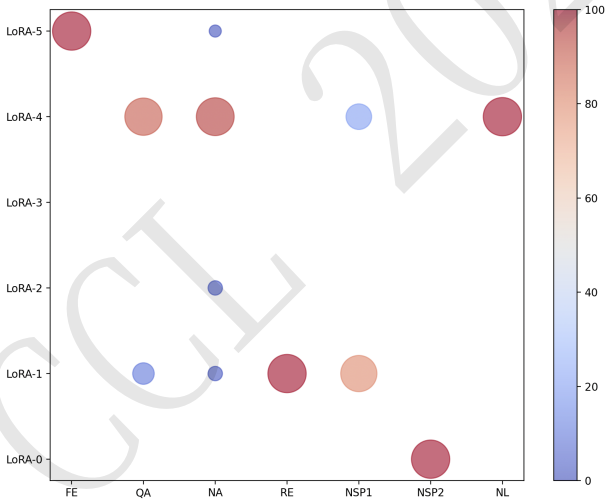



Figure 5: CFLEB数据集每类任务对专家的选择情况

### 4.5 消融实验

为了证明ASSL框架中算法的有效性，我们对以下三个问题进行探究：

1. 问题1：在语义空间进行聚类学习相较于人工任务分类是否存在优势?
2. 问题2：进行第一阶段数据重分布是否具有意义?
3. 问题3：进行第二阶段的数据筛选补充是否具有意义?

为了回答问题1，我们设计了一组对照实验，两个模型系统均采用LoRA专家自适应选择的方式。对于第一个模型系统，在数据预处理阶段，根据自然语言处理任务定义，人工将数据分为七类，每一类数据训练一个LoRA专家，组成多专家系统。同时，对于第二个模型系统，我

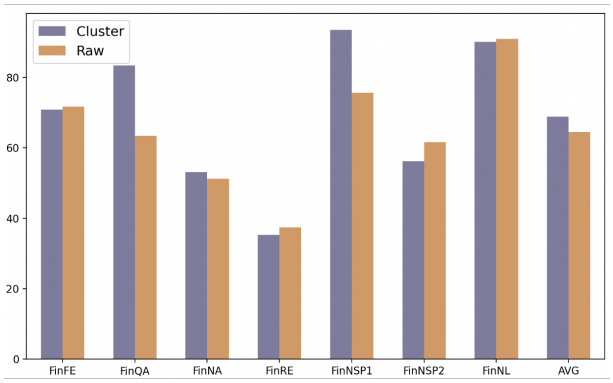



Figure 6: CFLEB数据集对于聚类和预定义模型系统的评测结果

们在语义空间中将数据重新聚类为六类，旨在验证语义空间中的距离能反映任务间的互补程度，相较于人工预定义类别具有优势。为此，我们对采用聚类算法得到的六类数据和原始的七类数据进行了详尽的对比分析，实验结果如 Figure 6 所示，预定义人工分类的方法在CFLEB上平均表现不如语义空间聚类的方法，在FinQA和FinNSP1任务上显著低于语义空间聚类方法。由此证明了语义空间聚类学习的方法相较于预定义任务类型方法的优越性，证明了利用语义空间相似性的方法能够聚拢互补任务的多种数据，起到促进模型在这些任务上的性能的作用。

为了回答问题2，我们将两种数据筛选算法作为基线：随机筛选以及K-Center贪婪算法，对ASSL框架中第一部分数据重分布的A-DBSCAN算法进行对比测试。三种数据筛选算法的侧重点不同，Random算法旨在实现从各聚类中按照原始数据分布情况均匀地抽取数据，K-Center贪婪算法注重选择数据的多样性，A-DBSCAN算法则对原始数据进行了重分布，根据密度的稀疏性对数据进行采样，同时也过滤了大量的离群噪声数据。实验结果如 Figure 7所示，在CFLEB数据集上，A-DBSCAN和Random算法表现突出；而在FinEval数据集上，K-Center贪婪和A-DBSCAN算法则展现了较好的性能。在两个数据集上都证实了A-DBSCAN算法的鲁棒性，表明这种通过对语义空间密度进行平滑分布方法的优越性，但由于其过滤了大量噪声，还需进一步进行数据补充。

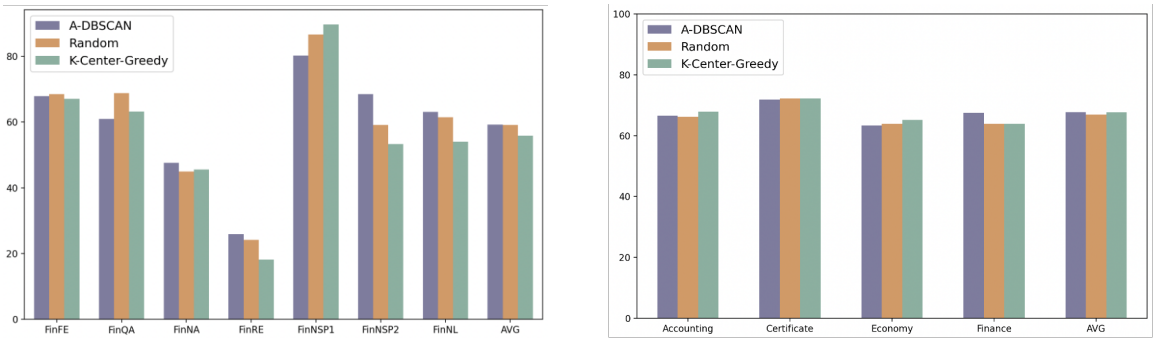
为了回答问题3，并进一步确认增强数据平滑分布的重要性，我们将利用效用函数筛选数据与Random算法筛选数据的实验结果进行对比，为保证公平性，两个实验类型均在每个聚类中选择4000条数据进行训练。实验结果如 Figure 8 所示，通过第二阶段的数据扩充，模型的综合性能相比于仅经过第一阶段筛选有了显著的提升，同时优于Random算法，这进一步验证了第二阶段进行数据扩充的必要性和有效性，增强了对于原始数据长尾分布的平滑性。

## 5 讨论

我们的研究基于设计ASSL框架训练了一个中文金融领域的多任务大模型，并在多个评测集上取得优异的性能。通过ASSL框架，研究人员可以深入了解基于语义空间的自适应学习优势，利用语义空间的特性，将多专家系统的优势注入语义空间中，实现专家自适应选择、数据自适应选择，从而得到更好的模型性能。基于此，我们可以进一步讨论这种基于语义空间自适应学习的互补特性。

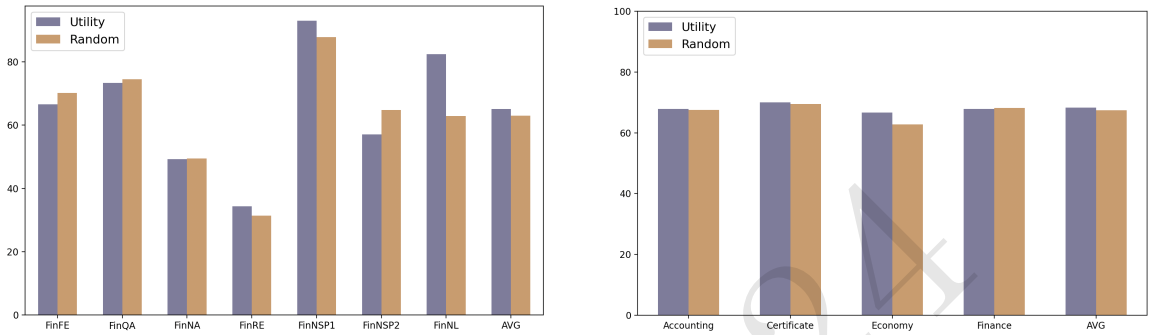
每个领域都拥有众多不同的自然语言处理或领域特定任务，要求领域大模型具备全面而综合的能力，但目前的多源异构数据容易引发模型进行任务风格迁移时的冲突。进行多任务学习时，利用所有语料对一个大模型进行训练极易影响模型指令遵循的能力，进而引发性能下降、幻觉产生等现象。基于此，制定一个多专家系统，并根据所有数据在同一个语义空间中的分布情况，为模型制定自适应的任务分类、数据筛选的方案变得尤为重要。根据 Figure 5，利用语义空间数据点的相似性能够有效的聚拢相互促进模型能力提升的任务数据，分隔互斥的任务数据，提高微调数据的均衡性。利用语义空间相似性不仅优化了数据分类的过程，也提供了一种高效的LoRA专家模型的挑选机制。通过比较问题嵌入与各个LoRA专家模型嵌入之间的相似性，系统能够精准地选取与特定问题最为匹配的LoRA专家模型。这种方法能够确保数据、任务与模型在语义空间上的连贯性与一致性，最终提高模型的综合能力。实验结果进一步验证了该方法的有效性和鲁棒性。相较于随机选择数据进行模型微调，采用数据筛选算法能够更加精


(c) 2024 中国中文信息学会计算语言学专业委员会




(a) CFLEB评估数据集　　(b) FinEval评估数据集

Figure 7: 消融实验2:第一阶段数据筛选算法对比




(a) CFLEB评估数据集　　(b) FinEval评估数据集

Figure 8: 消融实验3:第二阶段数据筛选算法对比

确地匹配合适的LoRA专家模型，显著提高了模型处理不同领域问题时的整体表现。

## 6　结论

本研究中我们主要探索了基于语义空间的自适应学习对于多专家大语言模型系统的性能影响，目标是根据语义空间对互补以及互斥的任务数据进行划分，利用其中的嵌入位置信息对多专家进行自适应的选择，并利用模型本身以及语义空间的密度分布对训练数据分布进行两阶段调整，该框架使得模型在进行多任务学习时具有较好的性能及泛化性。我们基于自适应语义空间学习框架，利用金融领域的公开数据集训练了“银瞳”大语言模型，并在基准上对“银瞳”模型系统进行了评测，显示出了其优异的表现。这一结果不仅证明了我们方法的可行性和有效性，也为未来多专家系统的大语言模型发展开辟了新的视野。

## 致谢

本研究得到了国家重点研发计划（2023YFC3304800）、中国工程院战略研究咨询项目“数字化转型背景下的金融风险监测与预警系统战略研究”（2023-XY-43）和上海市自然科学基金（23ZR1404900）的支持。

## 参考文献

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336.

Wei Chen, Qiushi Wang, Zefei Long, Xianyin Zhang, Zhongtian Lu, Bingxuan Li, Siyuan Wang, Jiarong Xu, Xiang Bai, Xuanjing Huang, et al. 2023. Disc-finllm: A chinese financial large language model based on multiple experts fine-tuning. *arXiv preprint arXiv:2310.15205*.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *arXiv preprint arXiv:2306.16092*.

Shihan Dou, Enyu Zhou, Yan Liu, Songyang Gao, Jun Zhao, Wei Shen, Yuhao Zhou, Zhiheng Xi, Xiao Wang, Xiaoran Fan, et al. 2023. Loramoe: Revolutionizing mixture of experts for maintaining world knowledge in language model alignment. *arXiv preprint arXiv:2312.09979*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Michael I Jordan and Robert A Jacobs. 1994. Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2):181–214.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Dakuan Lu, Hengkui Wu, Jiaqing Liang, Yipei Xu, Qianyu He, Yipeng Geng, Mengkun Han, Yingsi Xin, and Yanghua Xiao. 2023. Bbt-fin: Comprehensive construction of chinese financial domain pre-trained language model, corpus and benchmark. *arXiv preprint arXiv:2302.09432*.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2023. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*.

James MacQueen et al. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

OpenAI. 2023. Gpt-4 technical report.

Leif E Peterson. 2009. K-nearest neighbor. *Scholarpedia*, 4(2):1883.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10):1872–1897.

Mengjie Ren, Boxi Cao, Hongyu Lin, Cao Liu, Xianpei Han, Ke Zeng, Guanglu Wan, Xunliang Cai, and Le Sun. 2024. Learning or self-aligning? rethinking instruction fine-tuning.

Erich Schubert, Jörg Sander, Martin Ester, Hans Peter Kriegel, and Xiaowei Xu. 2017. Dbscan revisited, revisited: why and how you should (still) use dbscan. *ACM Transactions on Database Systems (TODS)*, 42(3):1–21.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ted Zadouri, Ahmet Üstün, Arash Ahmadian, Beyza Ermiş, Acyr Locatelli, and Sara Hooker. 2023. Pushing mixture of experts to the limit: Extremely parameter efficient moe for instruction tuning. *arXiv preprint arXiv:2309.05444*.

CCL 2024

Liwen Zhang, Weige Cai, Zhaowei Liu, Zhi Yang, Wei Dai, Yujie Liao, Qianru Qin, Yifei Li, Xingyu Liu, Zhiqiang Liu, et al. 2023. Fineval: A chinese financial domain knowledge evaluation benchmark for large language models. *arXiv preprint arXiv:2308.09975*.

Ziyu Zhao, Leilei Gan, Guoyin Wang, Wangchunshu Zhou, Hongxia Yang, Kun Kuang, and Fei Wu. 2024. Loraretriever: Input-aware lora retrieval and composition for mixed tasks in the wild. *arXiv preprint arXiv:2402.09997*.

Yuhang Zhou, Yuchen Ni, Xiang Liu, Jian Zhang, Sen Liu, Guangnan Ye, and Hongfeng Chai. 2024. Are large language models rational investors? *arXiv preprint arXiv:2402.12713*.

CCL 2024

# A 附录

## A.1 A-DBSCAN

**Algorithm 1** 简化且包含子簇处理的基于密度自适应A-DBSCAN算法

**Data:** 数据集$D$, 邻域大小$k$
**Result:** 形成的簇集合$Clusters$
初始化簇集合$Clusters = \emptyset$ 计算每个数据点$p$的局部密度$\rho_p$:

$$\rho_p = \frac{1}{\frac{1}{k}\sum_{j=1}^{k} d(p, p_j)}$$

根据$\rho$对$D$中的点进行排序，形成优先队列$Q$ 初始化$\varepsilon$和$MinPts_{\text{init}}$:

$$\varepsilon = \text{中位数}\{d(p, p_k)\} \quad \forall p \in D$$

$$MinPts_{\text{init}} = \frac{\varepsilon \times \rho_{\max}}{2}$$

**while** 从$Q$中选取队列首个点$p$ **do**
  **if** 未访问过$p$ **then**
    标记$p$为已访问 获取$p$的$\varepsilon$-邻域中的点$N_p$ **if** $|N_p| \geq MinPts_{init}$ **then**
      初始化一个新簇$cluster$并添加$p$ **for** 每个点$q \in N_p$ **do**
        **if** $q$未被访问 **then**
          标记$q$为已访问 获取$q$的$\varepsilon$-邻域中的点$N_q$ **if** $|N_q| \geq MinPts_{init}$ **then**
            将$N_q$中的点添加到$N_p$中
          **end**
        **end**
        **if** $q$不属于任何簇 **then**
          将$q$添加到$cluster$
        **end**
      **end**
      $Clusters = Clusters \cup \{cluster\}$ 更新$MinPts$:

$$MinPts = \max\left(2, \frac{\rho_p}{\rho_{\max}} \times MinPts_{\text{init}}\right)$$

    **end**
  **end**
**end**
处理子簇($Clusters$) **return** $Clusters$

**Procedure** 处理子簇$Clusters$**:**
  计算平均量$N_{avg} = \frac{\sum_i N_{C_i}}{|Clusters|}$，其中$N_{C_i}$是子簇$C_i$的数据量
  **foreach** 子簇$C_i \in Clusters$ **do**
    **if** $N_{C_i} > N_{avg}$ **then**
      采样: 从$C_i$中随机选择$N_{avg}$个数据点
  **end**
**return**

## A.2 金融NLP任务

在这一部分，我们将会对用于构造指令的金融NLP数据集进行更为详细的介绍。数据集涵盖了7类任务，包括情感分析、信息抽取、文本分类、文本生成、语义匹配、金融问答和金融考

CCL 2024

试选择题，具体统计数据如 Table 3所示。

Table 3: 数据集统计

| 数据集 | 主要任务类型 | 细分任务类型 | 数量（条） |
|---|---|---|---|
| flare-zh-fe | 情感分析 | 情感分析 | 2020 |
| flare-zh-stocka | 情感分析 | 情感分析 | 1477 |
| flare-zh-stockb | 情感分析 | 情感分析 | 1962 |
| CFSC-ABSA | 情感分析 | 情感分析 | 33184 |
| ficuge-finnl | 文本分类 | 金融新闻分类 | 7071 |
| flare-zh-nl | 文本分类 | 金融新闻分类 | 884 |
| flare-zh-nl2 | 文本分类 | 金融新闻分类 | 884 |
| flare-zh-nsp | 文本分类 | 金融负面消息及其主体判定 | 500 |
| fincuge-finna | 文本生成 | 新闻摘要 | 28799 |
| flare-zh-na | 文本生成 | 新闻摘要 | 3600 |
| flare-zh-19ccks | 信息抽取 | 事件类型与实体识别 | 2936 |
| flare-zh-20ccks | 信息抽取 | 事件类型与实体识别 | 9159 |
| flare-zh-21ccks | 信息抽取 | 事件类型与实体识别 | 1400 |
| flare-zh-22ccks | 信息抽取 | 事件类型与实体识别 | 11829 |
| fincuge-finre | 信息抽取 | 关系抽取 | 13486 |
| flare-zh-ner | 信息抽取 | 命名实体识别 | 337 |
| flare-zh-corpus | 语义匹配 | 语义匹配 | 10000 |
| flare-zh-afqmc | 语义匹配 | 语义匹配 | 4316 |
| fincuge-finqa | 金融问答 | 问答 | 19906 |
| fincuge-fincqa | 金融问答 | 因果问答 | 21965 |
| Duxiaoman/FinanceIQ | 金融考试选择题 | 单项选择题 | 3573 |
| fingpt-fineval | 金融考试选择题 | 单项选择题 | 1056 |
| Duxiaoman/FinCorpus | 金融考试选择题 | 多项选择题 | 40000（过滤后） |

**情感分析** 情感分析能够反映金融新闻的情感趋向，影响决策。我们采集了4个数据集，包括flare-zh-fe、flare-zh-stocka、flare-zh-stockb和CFSC-ABSA。flare-zh-stocka分析市场数据和公司相关的公告，根据新闻对股票数据的影响判断公司的股票运动走势是跑赢大盘、中性还是表现不佳。其余数据集是判断所给的金融新闻中句子的情绪，给出积极、消极或中性的回答。

**文本分类** 对于金融新闻分类，选取ficuge-finnl、flare-zh-nl、flare-zh-nl2这3个数据集，根据所给金融文本，输出2-3个类别关键词。对于金融负面消息及其主体判定，flare-zh-nsp数据集根据金融新闻和实体，判断该实体是否含有负面消息。

**文本生成** 文本生成任务主要是根据金融新闻长文本生成摘要，要求简洁但能包含关键信息，使用的数据集有fincuge-finna和flare-zh-na。

**信息抽取** 信息抽取任务主要包括事件类型与实体识别、关系抽取和命名实体识别，使用的数据集有flare-zh-19ccks、flare-zh-20ccks、flare-zh-21ccks、flare-zh-22ccks、fincuge-finre和flare-zh-ner。其中前四个数据集属于事件类型与实体识别类型，分析给定金融文本，判断其中所有的事件类型及其对应主体；关系抽取数据集fincuge-finre根据金融文本及实体对，从44个关系中选择出能正确表示该实体对的关系；命名实体识别数据集flare-zh-ner是对金融文本中存在实体的抽取。

**语义匹配** 对于语义匹配任务，我们选取了两个数据集：flare-zh-corpus和flare-zh-afqmc，判断两个金融文本表达的语义是否一致。

**金融问答** 针对普通的金融问答，使用fincuge-finqa数据集，从金融文本中提取事件信息，并结合上下文回答问题。而金融因果数据集fincuge-fincqa还关注金融文本中的因果关系。

**金融考试选择题** 包括来自Duxiaoman和FinGPT的金融各领域的单项选择题，以及Duxiaoman金融预料中的多项选择题，涵盖了银行、证券、会计、基金、期货等多领域的知识和计算。

第二十三届中国计算语言学大会论文集，第955页–第972页，太原，中国，2024年7月25日至28日。
卷1：主会论文
(c) 2024 中国中文信息学会计算语言学专业委员会

### A.3 金融NLP任务示例

将所有训练数据统一为三元组{instruction, input, output}格式，具体各类数据集的示例如图 Figure 9 - Figure 13所示。


```
{
    "instruction": "针对金融股民论坛的这段话，请评估它的情绪倾向：是积极、消极，还是处于中立态度？欧股涨幅扩大，德国DAX指数、欧洲斯托克50指数均涨超1%，法国CAC40指数涨0.82%，英国富时100指数涨0.71%。\n\n答案中只需包含\"积极\"、\"中性\"或\"消极\"",
    "input": "欧股涨幅扩大，德国DAX指数、欧洲斯托克50指数均涨超1%，法国CAC40指数涨0.82%，英国富时100指数涨0.71%。\n",
    "output": "积极"
}
```


Figure 9: 情感分析示例


```
{
    "instruction": "阅读以下金融声明，并根据其内容，从中国、外国、国际、公司、行业、大盘、经济、政策、政治、期货、债券、房地产、外汇、虚拟货币、新冠、能源中挑选两至三个最相关的类别，答案用空格隔开。\n国家外汇管理局：加强风险防控和监管能力建设方面，完善外汇市场“宏观审慎+微观监管”两位一体管理框架，加强跨境资金流动风险监测预警和逆周期调节，严厉打击外汇领域违法违规活动，维护外汇市场稳定和健康秩序。\n\n",
    "input": "国家外汇管理局：加强风险防控和监管能力建设方面，完善外汇市场“宏观审慎+微观监管”两位一体管理框架，加强跨境资金流动风险监测预警和逆周期调节，严厉打击外汇领域违法违规活动，维护外汇市场稳定和健康秩序。
",
    "output": "中国政策"
}
```


Figure 10: 文本分类示例


```
{
    "instruction": "\n 央行金融消费权益保护局副局长尹优平日前表示，人民银行将统筹推进违法金融广告和营销宣传行为治理，科学运用各种技术手段，有效抓取非法金融广告线索；加强金融矛盾源头预防、前端化解，为人民群众提供低成本、高效率、一站式、法制化的金融纠纷化解渠道；提高金融监管的灵敏度、精准度和及时性，推动治理能力现代化建设。\n\n阅读后，请简述所给资讯的核心要点，摘要长度限制在一到两句话之内。",
    "input": " 央行金融消费权益保护局副局长尹优平日前表示，人民银行将统筹推进违法金融广告和营销宣传行为治理，科学运用各种技术手段，有效抓取非法金融广告线索；加强金融矛盾源头预防、前端化解，为人民群众提供低成本、高效率、一站式、法制化的金融纠纷化解渠道；提高金融监管的灵敏度、精准度和及时性，推动治理能力现代化建设。",
    "output": "央行：将统筹推进违法金融广告和营销宣传行为治理"
}
```


Figure 11: 文本生成示例


```
{
    "instruction": "\n此后外资对国内OTC重磅品种的并购沉寂数年,直到日前拜耳宣布全资收购滇虹药业,OTC并购战火才重新燃起。头实体：拜耳，尾实体：滇虹药业。\n\n在财经金融相关的文本和问题中，只从提供的选项中挑选出一个正确答案，直接给出这个答案，不包含其他任何信息。你可以选择的关系类别包括：分析、借壳、合作、转让、买资、入股、商讨、被拟收购、被成立、重组、自己、被注资、被分析、被帮助、交易、被持股、被入股、注资、成立、被买资、被借壳、增持、被拥有、发行、订单、拥有、纠纷、被增持、被转让、合资、减持、欠款、其他、被减持、签约、拟收购、被收购、合并、帮助、被发行、被欠款、持股、收购、竞争\n请注意，如果头尾实体不含以上关系，请输出其他。",
    "input": "此后外资对国内OTC重磅品种的并购沉寂数年,直到日前拜耳宣布全资收购滇虹药业,OTC并购战火才重新燃起。头实体：拜耳，尾实体：滇虹药业。",
    "output": "收购"
}
```


Figure 12: 信息抽取示例

### A.4 数据混合比例分布


(c) 2024 中国中文信息学会计算语言学专业委员会

```
{
    "instruction": "分析两个金融语句的意思是否一致，你只需要回答是或否。请分析下面所给语句语义是否一致：\n",
    "input": "使用花呗人脸识别不了怎么办\n花呗要人脸识别 识别不了怎么回事",
    "output": "是"
}
```


Figure 13: 语义匹配示例


```
{
    "instruction": "\n\n上汽集团股份回购事件对应的回购股份数量是什么？上汽集团公告，公司计划回购5842万股-1.17亿股股份，占公司总股本的0.5%-1%，拟回购资金总额为不超过人民币30.34亿元，回购实施期限自股东大会审议通过本次回购方案之日起6个月内。\n\n针对本段上下文和提出的问题，提取信息并作答。如不能回答，请回答无相应参数。",
    "input": "上汽集团股份回购事件对应的回购股份数量是什么？上汽集团公告，公司计划回购5842万股-1.17亿股股份，占公司总股本的0.5%-1%，拟回购资金总额为不超过人民币30.34亿元，回购实施期限自股东大会审议通过本次回购方案之日起6个月内。",
    "output": "5842万股-1.17亿"
}
```


Figure 14: 金融问答示例


```
{
    "instruction": "请根据上下文题目，不论是单选题还是多选题，列出所有正确答案并进行分析。\n在个人/家庭资产负债表中，下列属于资产项目的有（）。I.现金与现金等价物II.股票、债券等金融资产III.车辆、房产等实物资产IV.信用卡透支\nA、I、II、IV\nB、I、II、III\nC、II、III\nD、I、III、IV\n",
    "input": "在个人/家庭资产负债表中，下列属于资产项目的有（）。I.现金与现金等价物II.股票、债券等金融资产III.车辆、房产等实物资产IV.信用卡透支\nA、I、II、IV\nB、I、II、III\nC、II、III\nD、I、III、IV",
    "output": "B\n解析：在个人/家庭资产负债表中，资产项目包括现金与现金等价物、股票、债券等金融资产、车辆、房产等实物资产。信用卡透支属于负债，因此不在资产项目中。选项B包含了正确的资产项目。
"
}
```


Figure 15: 金融考试选择题示例

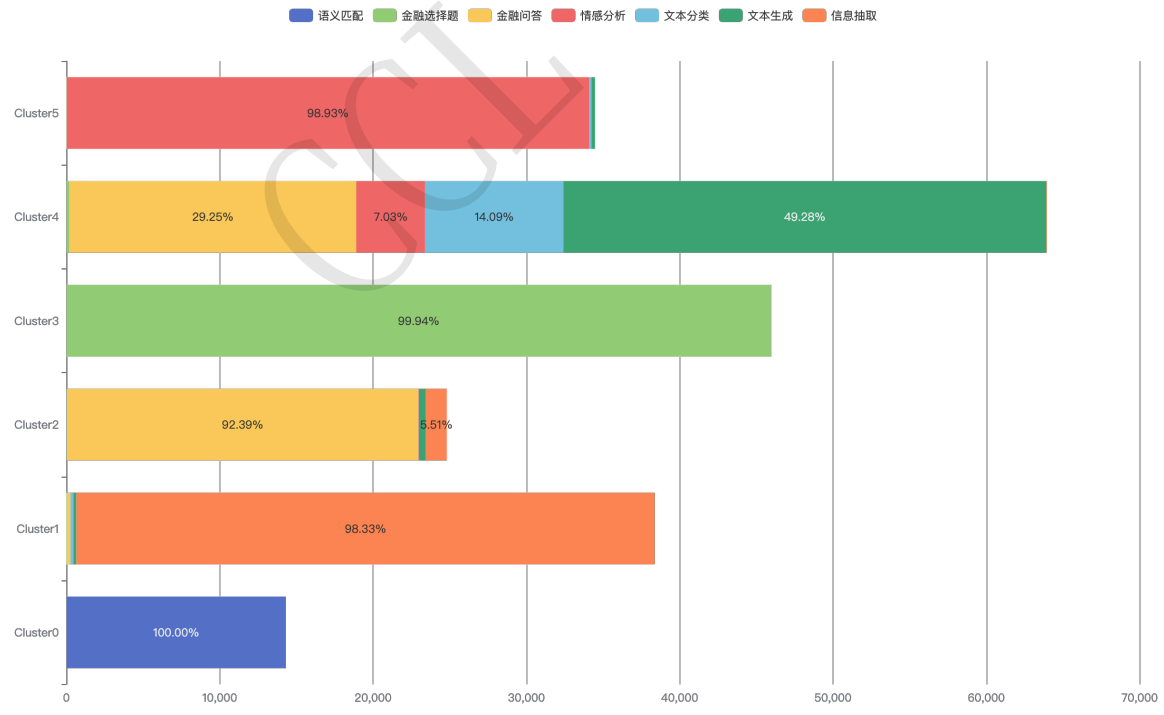



Figure 16: 经过聚类后每个类别的数据混合比例


(c) 2024 中国中文信息学会计算语言学专业委员会

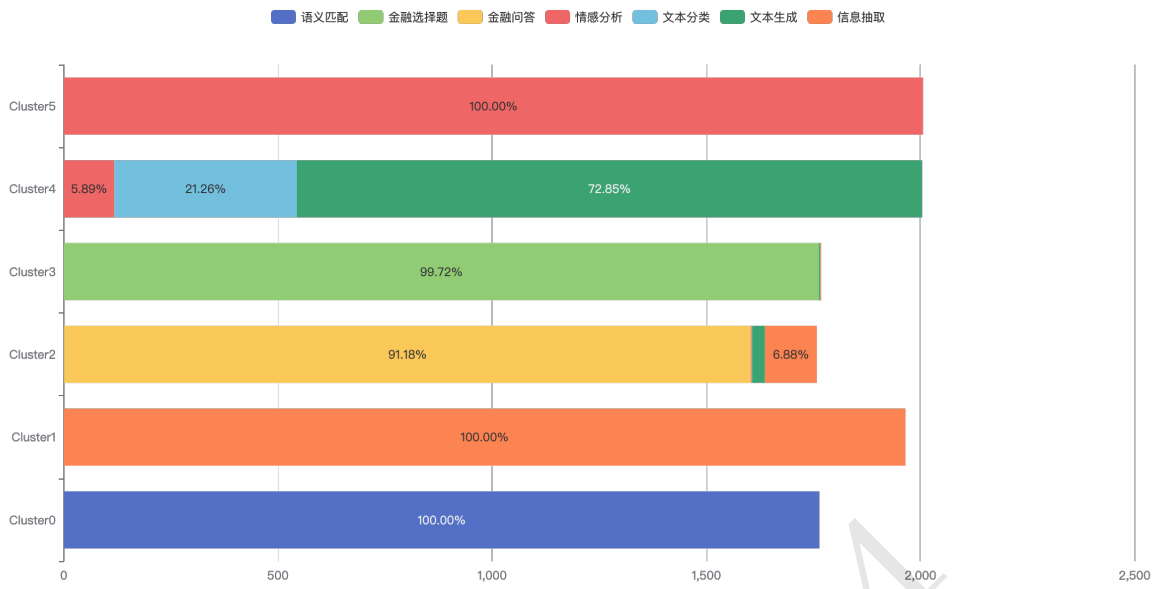



Figure 17: 经过一阶段数据重分布后，每个类别的数据混合比例

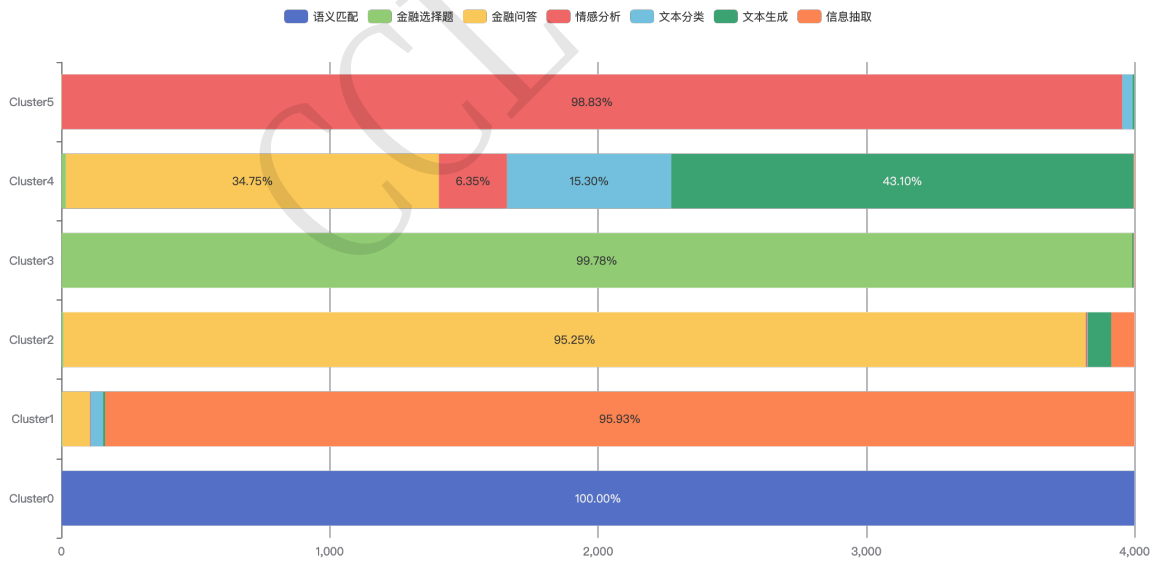



Figure 18: 经过二阶段数据重分布后，每个类别的数据混合比例

CCL 2024

# Enhancing Free-Form Table Question Answering Models by Distilling Relevant-Cell-Based Rationales

**Zhiyu Yang**[1] **Shuo Wang**[2] **Yukun Yan**[2] **Pengyuan Liu**[1,3]* **Dong Yu**[1]
[1]School of Information Science, Beijing Language and Culture University
[2]Department of Computer Science and Technology, Tsinghua University
[3]National Language Resources Monitoring and Research Center for Print Media
202121198419@stu.blcu.edu.cn
liupengyuan@pku.edu.cn

## Abstract

Free-form table question answering is a challenging task since tables contain structured contents compared to plain texts, which requires high-level reasoning abilities to effectively identify cells that are relevant to the question and produce a correct and faithful answer based on their relations. Large language models (LLMs) have exhibited remarkable reasoning capabilities in numerous NLP applications. However, in some specific tasks, specially-trained small models can still outperform LLMs. Furthermore, small models require extremely less computation costs compared to LLMs. To leverage the strengths of both types of models, we propose a **R**elevant-**C**ell-based **K**nowledge **D**istillation with inference-time **T**eacher **G**uidance **(RCKD-TG)** method. This approach aims to combine small free-form table question answering models' abilities to learn from human annotations and large language models' abilities to effectively reason from table contents, via applying Relevant-Cell-based rationales distilled from LLMs to small models' training and inference stages. Our experiments demonstrate the superiority of our method over vanilla small models in correctness, faithfulness, adequacy and fluency, also over general LLMs in adhering to the style of human annotations. We achieve state-of-the-art performance on FeTaQA, a representative free-form table question answering benchmark. Our result of a 41.3 BLEU score demonstrates the feasibility of effectively using small models' task-specific abilities and LLMs' reasoning capabilities at the same time. Additionally, our method exhibits high computation efficiency and data efficiency. Compared to strong baselines, we achieve better performance with significantly less training data.

## 1 Introduction

Tables are a prevalent form of structured data commonly found in databases and on the internet. Traditionally, tables were grounded using semantic parsing or by converting natural language queries into SQL language to obtain desired outputs. Representative benchmarks, like WTQ (Pasupat and Liang, 2015), SQA (Iyyer et al., 2017), WikiSQL (Zhong et al., 2017), Spider (Yu et al., 2018), TabFact (Chen et al., 2019), have been established for evaluating table grounding methods. Recent efforts have leveraged pre-training methods for table grounding tasks, with models like TAPAS (Herzig et al., 2020), TaBERT (Yin et al., 2020), TAPEX (Liu et al., 2021), and TableFormer (Yang et al., 2022) achieving promising results on existing benchmarks. However, these benchmarks primarily focus on providing short-form answers to queries, which are not challenging enough for models to perform complex reasoning.

With the emergence of large generative models, users increasingly prefer interacting with NLP systems through natural language queries and receiving responses in free-form answers (Chung et al., 2022; Ouyang et al., 2022; Thoppilan et al., 2022). As a result, traditional short-form table parsing benchmarks and methods fall short of meeting modern users' demands. To address this limitation, (Nan et al., 2021) introduced FeTaQA, a free-form table question answering benchmark, aiming to bridge the gap. FeTaQA challenges models to retrieve multiple relevant entities and reason over them to produce correct and

*Corresponding author.
©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 973-985, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

faithful answers given more abstract and ambiguous queries. This benchmark is thus more challenging than previous short-form counterparts.

Researchers have continued to explore table pre-training methods to tackle free-form table question answering. For example, GenTaP (Shi et al., 2022) leverages large-scale synthesized training data with more than 500K training samples. However, such methods may not fully enhance a model's reasoning capabilities. Recently, Large Language Models (LLMs) have exhibited remarkable reasoning capabilities across various NLP tasks. (Brown et al., 2020; Wei et al., 2022) have demonstrated LLMs' ability to quickly generalize to unseen tasks and perform reasonably well given only a few prompts. However, developing and deploying LLMs for each specific scenario is computationally prohibitive. To address this challenge, the latest works (Li et al., 2023; Hsieh et al., 2023) have adopted Knowledge Distillation, using LLMs as Teacher models and small models as Student models. These methods train small models on LLMs' output to distill their knowledge and reasoning capabilities, and in the meantime, outperform their Teacher models. While successful in classification and multiple-choice QA tasks, their effectiveness in generation tasks and structured data remains rather unexplored.

It is harder to distill knowledge from larger teacher models in text generation settings because of inadequate supervision signal. Automatic evaluation metrics for language generation tasks are not robust enough to assess all important aspects pertinent to the quality of generated text. They succeed in evaluating the surface similarities between the generated text and the reference text, but fails in evaluating the accuracy, faithfulness and fluency of generated text. Therefore, training a small model using these metrics as supervision would constrain to the small model's output to be lexically and syntactically more similar to human annotations, but unable to guide the small model to generate semantically correct and grammatically fluent answers. Meanwhile, LLMs could provide mostly correct answers but fail to conform to human annotated preferences, and it is computationally prohibitive to fine-tune LLMs on human annotated data to mitigate this problem.

In light of this, we propose a simple method for free-form table question answering, **R**elevant-**C**ell-based **K**nowledge **D**istillation with inference-time **T**eacher **G**uidance **(RCKD-TG)**, to tackle the challenges of reasoning on table contents and generating free-from answers that are more preferred towards human annotations. We devise a specific training method, so that at inference time, LLM's abilities to reason and small model's abilities to produce answers more similar to human annotations can be combined effectively.

Firstly, on the training data, we prompt the LLM to generate information in cells that are relevant to answering the question. Then we use these rationales as references to teach the small model to produce crucial information autonomously. A multi-task learning framework is employed, training the small model on both the LLM's output and the QA data. Secondly, during inference, we augment the Student model's generation process with LLM's Relevant-Cell-based rationale, yielding results surpassing those using small models or LLMs alone. With less than 10K training data, our method achieves a state-of-the-art 41.3 BLEU score on FeTaQA, demonstrating the feasibility of combining LLMs' powerful reasoning and small models' task-specific capabilities while maintaining high data and computation efficiency compared to previous table pre-training methods.

In summary, the contributions of this paper are:

- Proposing Relevant-Cell-based Knowledge Distillation with inference-time Teacher Guidance to combine small models' capabilities of learning from human annotation and LLMs' capabilities of effectively reasoning and providing correct answers.
- Achieving state-of-the-art performance on FeTaQA through our proposed method.
- Demonstrating the feasibility of combining LLMs' reasoning strength and specially-trained small models' task-specific strength in a challenging downstream task.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 973-985, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

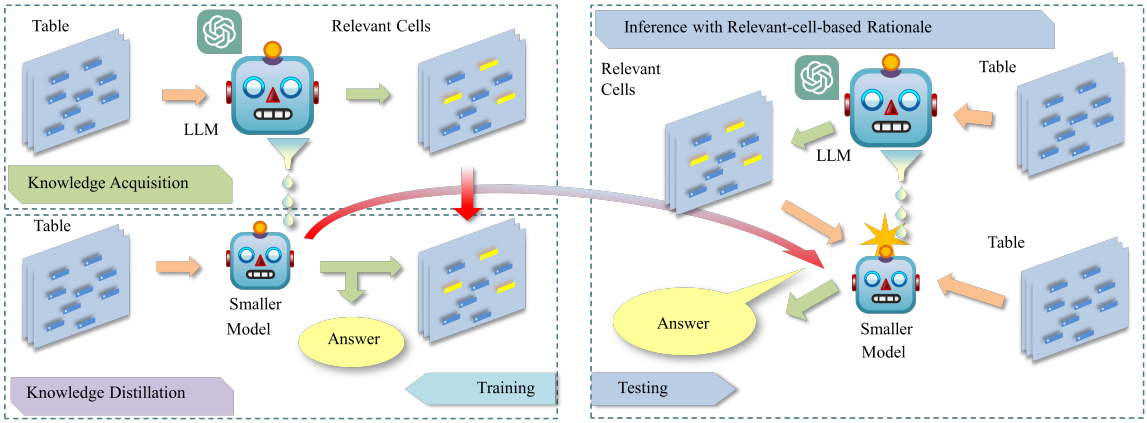



Figure 1: Overview of our proposed method. In Knowledge Acquisition phase, we prompt LLM to produce Relevant-Cell-based Rationales. In Knowledge Distillation phase, we distill LLM's knowledge to the Student model by using Relevant-Cell-based Rationales as target sequences for Student to generate. In the meantime, the Student model is trained on the table question answering data, forming a multi-task learning framework. At inference phase, we found that incorporating Teacher model's rationales as guidance for Student model can greatly improve the quality of Student model's answers.

## 2 Related Work

### 2.1 Table-to-Text Generation

To combine Natural Language Generation and table grounding, prior research has largely focused on the domain of table-to-text generation (Chen et al., 2020; Parikh et al., 2020; Cheng et al., 2021; Liu et al., 2022b; Nan et al., 2021; Zhao et al., 2023b). These works are more closely related to controllable text generation rather than question answering, as they often involve generating correct and faithful descriptions using highlighted table regions as controlling elements. For instance, datasets like ToTTo (Parikh et al., 2020) and HiTab (Cheng et al., 2021) require models to convert highlighted table segments into single-sentence summaries. LogicNLG (Chen et al., 2020) revolves around generating statements that can be logically derived from facts within provided table regions. RotoWire (Lebret et al., 2016) requires models to craft summaries based on basketball game tables, while SciGen (Moosavi et al., 2021) and NumericNLG (Suadaa et al., 2021) datasets demand arithmetic reasoning and description generation for tables from scientific papers. QTSumm (Zhao et al., 2023c) tasks models with generating comprehensive summaries encompassing key information from tables spanning diverse topics.

### 2.2 Table Question Answering

Distinct from table-to-text generation, Table QA involves responding to user queries based on information from a source table. Notable datasets such as WikiTableQuestions (Pasupat and Liang, 2015), SQA (Iyyer et al., 2017), and WikiSQL (Zhong et al., 2017) primarily emphasize short-form answers. To address these Table QA tasks, researchers have employed table-based pretraining techniques (Herzig et al., 2020; Eisenschlos et al., 2020; Shi et al., 2022; Yin et al., 2020; Liu et al., 2021). A more recent addition, FeTaQA (Nan et al., 2021), adopts ToTTo's (Parikh et al., 2020) statements, reformulating them into questions and utilizing the same statements as answers. In this work, we select FeTaQA as the benchmark to evaluate out methods.

### 2.3 Knowledge Distillation via LLMs

Ever since (Hinton et al., 2015) proposed the Knowledge Distillation method, many works have followed its white-box distillation paradigm. Notably, recent approaches have taken up symbolic knowledge distillation (West et al., 2021), also referred to as black-box knowledge distillation or sample matching.

In contrast to distilling soft representations such as logits from larger models, this approach employs a black-box model, often an LLM, to generate training samples for Student models (Petroni et al., 2019; West et al., 2021; Liu et al., 2022a; Bhagavatula et al., 2022). Recent progress in LLM knowledge distillation includes the works of (LI et al., 2022; Huang et al., 2022; Magister et al., 2022), and (Ho et al., 2022), all demonstrates that small models can gain insight from the Chain-of-Thought (CoT) rationales of larger models. However, these works focus on reasoning tasks that yield definite answers, such as answer choices or numerical values. In contrast, our work focuses on free-form question answering, where responses are in the form of natural language text.

## 3 Methodology

### 3.1 Relevant-Cell-based Rationale Generation

Existing LLM Knowledge Distillation method does not perform ideally on TableQA tasks. Existing methods rely on CoT prompting to distill knowledge from LLMs. Such knowledge fails to produce detailed and crucial information. We proposed a novel and effective method to prompt LLM to generate useful knowledge for TableQA task (see Figure 1, upper left).

Prior works on LLM Knowledge Distillation predominantly focus on reasoning tasks where the input consists solely of questions (Li et al., 2023; Hsieh et al., 2023; Wang et al., 2022). These tasks require models to answer questions using knowledge incorporated into their parameters during pre-training, or acquired from external sources such as knowledge graphs or databases. However, for Table QA tasks, all necessary information to answer questions is presented within the input tables. Traditional CoT prompting method fails to adequately ground table content as well as its structured information, as the knowledge is too vague and abstract for small models to effectively learn from. To bridge this gap, we introduce a novel prompting technique designed specifically for extracting Relevant-Cell-based rationales.

Past prompt templates typically follow (Wei et al., 2022)'s format, with which LLMs are prompted to generate step-by-step reasoning sequences given a question and answer choice. These templates consist of question $x_i$, answer choice $y_i$, and human-written rationales $z_i$. However, our task of Free-Form Table Question Answering involves questions and tables as inputs, free-form sentence answers as output. Consequently, we extend the template to

$$\mathcal{P} = \{x_p, t_p, y_p, r_p\} \tag{1}$$

where $x_p$ represents questions, $t_p$ represents tables, $y_p$ denotes answer references, and $r_p$ represents our hand-crafted exemplars for extracting Relevant-Cell-based rationales.

To construct our prompt template, we randomly select ten examples $(x_p, t_p, y_p)$ from the training set

$$\mathcal{D}_{train} = \left\{\left(x_i^{train}, t_i^{train}, y_i^{train}\right)\right\}_{i=1}^{|D_{train}|} \tag{2}$$

and manually craft Relevant-Cell-based rationale $r_p$ exemplars for the selected examples. These exemplars contain reasoning paths from question $x_i$ to relevant table cells, coupled with explicit mention of the cells' corresponding column header names.

We argue that including these explicit column headers aids the small model in bridging the semantic gap between the question and cell content, thereby enhancing its ability to locate relevant information in the tables. Furthermore, these column headers denote the spatial positioning of table cells, acting as implicit guidance for the small model to comprehend table structural information. This collection of ten exemplars forms the seed set $\mathcal{P}$, with one exemplar $\mathcal{P}_k = (x_k, t_k, y_k, r_k) \sim \mathcal{P}$ randomly selected each time the LLM is prompted. This selected exemplar $\mathcal{P}_k$ is prepended to the question $x_i$ and table input $t_i$, then the LLM generate the Relevant-Cell-based rationale $r_i$, and this process is repeated across the entire dataset, thus equipping each example with its distinct Relevant-Cell-based rationale. The result of this process is a corpus

$$\mathcal{C}_{train} = \{(x_i, t_i, y_i, r_i)\}_{i=1}^{|D_{train}|} \tag{3}$$

which contains training data and its corresponding rationales.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 973-985, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

### 3.2 Multi-Task Learning Framework Knowledge Distillation

In order to effectively utilize LLM generated Relevant-Cell-based rationales in both the training stage and inference stage, the following training framework is proposed.

There are two ways of incorporating rationales into the training process. The first is to put it in the input, and teach the small model to generate an answer given the question, table and rationale. $f(x_i, t_i, r_i) \rightarrow y_i$, where $f$ denotes the small model.

The second way is to use the rationale as the target sequence, and ask the small model to generate it when given the question and table: $f(x_i, t_i) \rightarrow r_i$.

We argue that, in order to improve the small model's capabilities to reason and extract crucial information from structured data, it is imperative to use the rationale as the target sequence. In this way, the small model can effectively learn how to generate Relevant-Cell-based rationales, so that it can obtain the ability to perform appropriate reasoning steps when given a table and the question.

So that in the training phase, there are two target sequences for the small model, one is the Relevant-Cell-based rationale, the other is the answer to the question. While using the answers as the target sequences, the Relevant-Cell-based rationales are appended to the tables and questions to form the complete input sequences. By doing this, the small model can be trained to adopt Teacher's guidance at test time, as it has been familiarized with this input-output mapping during training.

So, in one optimizing step, the model is required to output model generated rationale and answer. Two separate losses

$$\mathcal{L}_{rationale} = \frac{1}{|C_{train}|} \sum_{i=1}^{|C_{train}|} l(f(x_i, t_i), r_i) \tag{4}$$

$$\mathcal{L}_{answer} = \frac{1}{|C_{train}|} \sum_{i=1}^{|C_{train}|} l(f(x_i, t_i, r_i), y_i) \tag{5}$$

are calculated between model generation and their respective ground-truth references. The total loss of

$$\mathcal{L} = \mathcal{L}_{rationale} + \mathcal{L}_{answer} \tag{6}$$

is then calculated. Consequently, a multi-task learning framework is formed (see Figure 1, lower left).

Additionally, we followed Distilling step-by-step and prepended task prefixes to the input of two different tasks. We prepended `[Relevant cells]` to the input to indicate the small model to generate Relevant-Cell-based rationale, and prepended `[Answer]` to indicate the small model to generate an answer.

### 3.3 Student Model Inference with Teacher Guidance

Using the multi-task learning framework, the small model alone can now be finetuned to produce answers that are similar to the surface forms of human annotation, but its ability to produce factually accurate, faithful and adequate answers remains limited. We propose a simple test-time augmentation to address this challenge. First, the Teacher model is prompted on the test set to produce Relevant-Cell-based rationales $r_{teacher}$ for the small model, serving as Teacher Guidance. When using the small model to generate answers $\hat{a}$ on the test set, the Teacher Guidance is appended to the tables $t_i$ and questions $x_i$ as input for the small model. We call this method **Inference with Teacher Guidance** (see Figure 1, right). The formulation for it is: $f(x_i, t_i, r_{teacher}) \rightarrow \hat{a}$, where $f$ denotes the finetuned small model.

We also tested small model's performance to produce answers on its own, without using Teacher's Relevant-Cell-based rationales at test time. We let the finetuned small model directly generate answer $\hat{a}$ given the question $x_i$ and table $t_i$ in a test example. We call this method **Direct Inference**. This setting helps us better understand the effectiveness of our proposed Relevant-Cell-based Rationales compared to other baselines that does not involve the Teacher model at inference time.

CCL 2024

| Model | sacreBLEU | ROUGE-1 | ROUGE-2 | ROUGE-L | METEOR |
|---|---|---|---|---|---|
| T5-large (Nan et al., 2021) | 30.5 | 0.63 | 0.41 | 0.53 | 0.49 |
| TAPEX (Liu et al., 2021) | 30.2 | 0.62 | 0.40 | 0.51 | - |
| ReasTAP (Zhao et al., 2022) | 30.4 | 0.63 | 0.40 | 0.51 | - |
| UnifiedSKG (Xie et al., 2022) | 32.4 | 0.64 | 0.42 | 0.54 | 0.51 |
| GenTaP (Shi et al., 2022) | 36.7 | 0.69 | 0.48 | 0.59 | 0.55 |
| GPT-3.5-turbo (1-shot) | 21.9 | 0.60 | 0.37 | 0.49 | 0.55 |
| Ours, trained w/ CoT, **Direct Inference** | 34.1 | 0.65 | 0.44 | 0.56 | 0.52 |
| Ours, trained w/ RC, **Direct Inference** | 34.3 | 0.66 | 0.44 | 0.56 | 0.52 |
| Ours, trained w/ CoT, **Inference with Teacher Guidance** | 31.5 | 0.63 | 0.41 | 0.53 | 0.50 |
| Ours, trained w/ RC, **Inference with Teacher Guidance** | **41.3** | **0.70** | **0.50** | **0.61** | **0.58** |

Table 1: Main results of our method comparing to baselines on the FeTaQA benchmark. CoT stands for Chain-of-Thought rationales. RC stands for Relevant-Cell-based Rationales. Due to the limitation of context window, only one dataset example can be fitted into the prompt template for GPT-3.5-turbo.

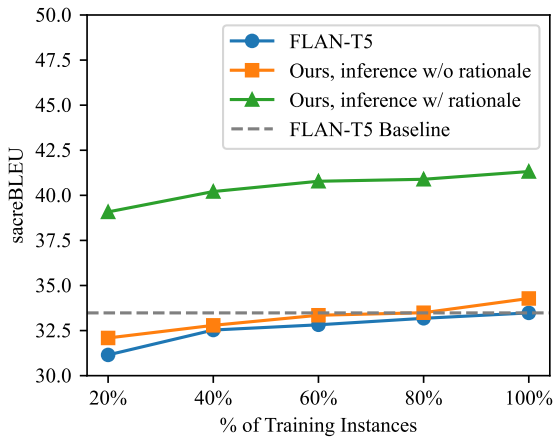



Figure 2: sacreBLEU score on FeTaQA benchmark with different amounts of training data.

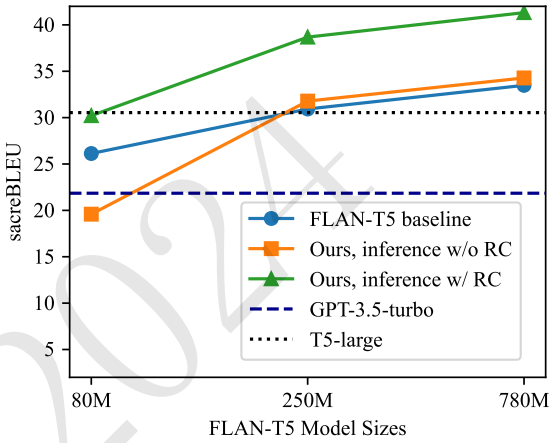



Figure 3: sacreBLEU score on FeTaQA benchmark with different sizes of FLAN-T5 model.

## 4 Experiments

In the upcoming experiments, we present a comprehensive analysis of our proposed **RCKD-TG** method. Firstly, we showcase the impressive performance of our method in comparison to previous works and the current state-of-the-art, highlighting its competitiveness across various evaluation metrics. Secondly, we explore the impact of dataset scale and model size, demonstrating our approach's remarkable data and computational efficiency. Thirdly, we conduct an in-depth evaluation of the answers generated by our method compared to those produced by LLMs. Also, we assess the quality of the Relevant-Cell-based rationales generated by both Teacher and Student model. Lastly, we delve into the capabilities of instruction-tuned models to effectively follow rationales and explore the impact of different types of instructions.

### 4.1 Datasets

We evaluate the performance of our method using the FeTaQA (Nan et al., 2021) benchmark, a comprehensive free-form table question answering benchmark comprising more than 10,000 (question, table, free-form answer, highlighted table cells) data samples. Unlike previous studies on Table Question Answering (Pasupat and Liang, 2015; Zhong et al., 2017; Yu et al., 2018) and LLM Knowledge Distillation (Li et al., 2023; Hsieh et al., 2023; Wang et al., 2022), FeTaQA presents unique challenges. The questions within this benchmark exhibit greater complexity, requiring intricate reasoning across multiple table cells. Additionally, the answers are expressed in a free-form manner, which demands the generation of

| Method | sacreBLEU |
|---|---|
| Ours, **Direct inference** | 34.3 |
| Ours, **Inference with Teacher Guidance** | 41.3 |
| GPT-3.5-turbo, 0-shot | 21.6 |
| GPT-3.5-turbo, 1-shot | 21.9 |
| GPT-3.5-turbo, w/ surface form constraint | 24.1 |

Table 2: Automatic evaluation results of GPT-3.5-turbo and our method's predicted answers.

in-depth elaborations and explanations as model predictions.

### 4.2 Implementation Details

In our implementation, the input is a pair of a question and its corresponding table. We followed UnifiedSKG to flatten the table as $\mathcal{T} = col : h, row_1 : r_1, ..., row_n : r_n$, where $h$ is column header, $r_i$ is the $i-th$ table row, "$col$" and "$row$" are natural texts indicating the region of table column headers and rows respectively. We also use a vertical bar "|" to distinguish between cells. The sequence begins with "`Table:` ", indicating the forthcoming linearized table information.

For generating Relevant-Cell-based rationales with LLM, we used GPT-3.5-turbo as our Teacher model. We sample from GPT-3.5-turbo with a one-shot prompt described in Methodology section and a temperature of 1.0. We sample one rationale from each data instance in FeTaQA training set and test set. For finetuning the small model on FeTaQA, we either use 8 NVIDIA RTX 3090 GPUs or 4 NVIDIA A40 GPUs. Initial learning rate is set to 5e-5 and a linear learning rate decay scheduler is employed. We use AdaFactor as the optimizer. We ran 40 epochs with the batch size of 64, and use early stopping with a patience of 5. The best checkpoint is selected according to the sacreBLEU score on the validation set.

### 4.3 Baselines

We conduct a comprehensive comparison of our method with several established baselines:

TAPEX (Liu et al., 2021): This method involves continuing pre-training of the BART model using synthetic SQL query execution data, enhancing model's table reasoning capabilities through SQL execution.

ReasTAP (Zhao et al., 2022): Based on the BART model, ReasTAP focuses on improving table understanding and reasoning abilities through pre-training on a synthetic corpus with questions that requires numerous types of reasoning skills to solve.

GPT-3.5-turbo: We perform one-shot testing on FeTaQA using our Teacher model alone and compare its performance with our proposed methods.

UnifiedSKG (Xie et al., 2022): This baseline unifies various structured knowledge grounding tasks with text-to-text language models, achieving remarkable performance across a range of table-to-text tasks.

It's noteworthy that ReasTAP and TAPEX did not originally evaluate their method on FeTaQA, we report their results from those published in this recent paper (Zhao et al., 2023a).

### 4.4 Models

Our standard Student model is FLAN-T5-large, which has 780M parameters. FLAN-T5 enhances T5 by scaling instruction fine-tuning and demonstrates better human-like reasoning abilities than the T5 model. We choose FLAN-T5 because it exhibited better rationale learning abilities compared with raw T5, as demonstrated in the following sections.

### 4.5 Main Results

Our primary results are obtained on the FeTaQA benchmark through training FLAN-T5-large Student model with Relevant-Cell-based rationales generated by GPT-3.5-turbo Teacher model. We also discuss the performance using regular CoT rationales during training, as shown in Table 1.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 973-985, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

| |
|---|
| **Ground Truth:** Janice Hahn received 33.3% of the vote against Newsom's 55.5%. |
| **Ours w/ CoT:** Janice Hahn performed less favorably compared to Gavin Newsom in terms of both votes and percentage of votes.<br>**Ours w/ RC:** Janice Hahn received 33.3% of the vote against Gavin Newsom's 55.5%. |

Table 3: An example of the CoT-trained *Inference with Teacher Guidance* prediction that fails to provide a detailed answer.

| |
|---|
| **Ground Truth:** At the World Championships, Winnie Ng participated twice in the marathon finishing 23rd in 1993 and 30th in 1995. |
| **T5-large:** At the 1995 World Championships in Gothenburg, Sweden, Winnie Ng finished 30th with a time of 3:01:08.<br>**Direct Inference:** Winnie Ng finished 23rd in the marathon at the 1993 World Championships and 30th at the 1995 World Championships.<br>**Teacher Guidance:** Winnie Ng participated in the World Championships twice in the marathon event, finishing 23rd in 1993 and 30th in 1995. |

Table 4: An example of our method's effectiveness on FeTaQA benchmark.

We report results from two testing modes: **Direct Inference** and **Inference with Teacher Guidance**. **Direct Inference** achieves a sacreBLEU score of 34.3, surpassing all methods except the existing state-of-the-art (SOTA), and outperforming the CoT rationale-trained model of 34.1 sacreBLEU score. **Inference with Teacher Guidance** achieves a sacreBLEU score of 41.32, establishing a new SOTA, while massively outperforming its CoT-trained counterpart which achieved 31.46. Using CoT-trained **Inference with Teacher Guidance** method could lead to small models generating vague and general answers, failing to point out specific details needed to thoroughly and correctly explain a query, as shown in Table 3.

In Table 2, we used 0-shot, 1-shot and 1-shot with surface form constraint settings to evaluate GPT-3.5-turbo on FeTaQA. For 1-shot with surface form constraint setting, we explicitly ask GPT-3.5-turbo to conform to the style of human annotated answer in the 1-shot example using detailed description. This method achieved a 24.1 sacreBLEU score compared to 0-shot setting's 21.6 and 1-shot setting's 21.9, showing marginal improvements as the efficacy of learning from in-context examples is limited.

Moreover, our method accomplishes this with just 7,326 training samples, along with an additional 7,326 rationales. Compared to previous SOTA's 500K synthetic pre-training corpus, our method demonstrates exceptional data efficiency. Furthermore, our approach outperforms the T5-3B model reported in UnifiedSKG, while utilizing a much smaller FLAN-T5 model with only 780M parameters, illustrating its computational efficiency.

### 4.6 Dataset Size Scaling and Model Size Scaling

We conducted experiments on datasets of varying proportions to investigate our method's performance in low-data regime. Training was performed on datasets comprising 20%, 40%, 60%, 80%, and 100% of the total data, as shown in Figure 2. Our Relevant-Cell-based rationale approach consistently outperforms the FLAN-T5 baseline. When using **Inference with Teacher Guidance**, our method trained on different

CCL 2024

| Source | Fluent | Correct | Adequate | Faithful |
|---|---|---|---|---|
| T5-large | 85.5 | 60.3 | 71.8 | 69.5 |
| Ours, trained w/ RC, **Direct inference** | 88.4 | 67.3 | 76.9 | 79.7 |
| Ours, trained w/ RC, **Inference with Teacher Guidance** | 91.3 | 74.1 | 79.7 | 85.9 |
| GPT-3.5-turbo | 93.7 | 80.1 | 83.5 | 87.7 |

Table 5: Human evaluation results of T5-large baseline, our method and GPT-3.5-turbo's predicted answers on FeTaQA.

| Source | Correct | Adequate | Faithful |
|---|---|---|---|
| LLM<br>Generated Knowledge | 84.3 | 81.3 | 97.9 |
| FLAN-T5<br>Generated Knowledge | 65.4 | 67.4 | 88.5 |

Table 6: Quality of Relevant-Cell-based rationales generated by either Teacher (LLM) or Student (FLAN-T5).

| Method | sacreBLEU | |
|---|---|---|
| | result | Δ |
| T5 | 32.5 | - |
| T5 w/ CoT | 32.5 | +0.0 |
| T5 w/ Relevant Cells | 33.4 | +0.9 |
| FLAN-T5 | 33.5 | - |
| FLAN-T5 w/ CoT | 34.1 | +0.6 |
| FLAN-T5 w/ Relevant Cells | 34.3 | +0.8 |

Table 7: Performance difference of T5 and FLAN-T5 trained on knowledge distilled from Teacher model and their respective baselines trained only on QA data.

dataset proportions all outperform the existing SOTA, which was achieved with 500K training samples, showcasing our method's data efficiency.

Similarly, we experimented with FLAN-T5 models with different sizes of full dataset to evaluate our method's effectiveness across model sizes, as shown in Figure 3. We found that our approach consistently surpasses the respective baseline of each model size, except for the 80M FLAN-T5-small model, where our method's performance drops. This phenomenon may be attributed to the small model's limited capacity to accommodate the complexity of two-task multi-task training. However, the performance significantly improves when using **Inference with Teacher Guidance**. This illustrates that even in situations where small models struggle to answer questions independently, introducing Teacher model knowledge can greatly enhance the small model's performance. This demonstrates the effectiveness of our combined small and large model approach.

### 4.7 Human Evaluation of Predictions and Rationales

We conducted comprehensive human and automatic evaluations of model predictions and Relevant-Cell-based rationales generated by the Student and Teacher models. We followed the human evaluation method in (Nan et al., 2021), and evaluated T5-large baseline, our proposed method and GPT-3.5-turbo's predictions in terms of Fluency, Correctness, Adequacy, and Faithfulness. Three NLP practitioners scored 50 model predictions on a scale of 0-5, which was then converted to a percentage scale, as shown in Table 5. Our methods consistently outperformed the T5-large baseline. **Inference with Teacher Guidance** method also exhibited notable gains in performance compared to **Direct Inference** method. Exemplified by Table 4, our methods exhibit significant improvements in answer completeness and correctness. Our best method achieved 97.3%, 92.5%, 95.4% and 97.9% of GPT-3.5-turbo's performance in terms of fluency, correctness, adequacy and faithfulness, while their automatic evaluation scores significantly surpassed those of GPT-3.5-turbo. This result demonstrates the feasibility of utilizing small model's ability of generating human annotated style text while retaining LLM's strong reasoning capabilities.

Similarly, we conducted human evaluations of the quality of Relevant-Cell-based rationales generated

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 973-985, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

| Instruction Type | # of instructions | | | | | |
|---|---|---|---|---|---|---|
| | 0 | 1 | 5 | 10 | 15 | 20 |
| No Instruction | 28.1 | - | - | - | - | - |
| Task Prefix | - | **34.3** | - | - | - | - |
| NL Instructions | - | 33.9 | 33.8 | 34.1 | 34.0 | 33.6 |

Table 8: Effect of different types and number of instructions on final performance on FeTaQA benchmark.

by FLAN-T5-large and GPT-3.5-turbo, as shown in Table 6. Three NLP practitioners evaluate 50 rationales based on three aspects: Correctness, Adequacy and Faithfulness. Our evaluations revealed that GPT-3.5-turbo's rationales significantly outperformed FLAN-T5-large's rationales in all aspects. Integrating FLAN-T5-generated rationales into the inference stage led to a decrease in performance, even lower than the baseline. This suggests that FLAN-T5's ability to self-generate rationales is limited, and directly using them to augment FLAN-T5's own inference stage hurts performance, demonstrating a disparity from previous works on classification and multiple choice tasks

### 4.8 Comparison of Instruction-Tuned FLAN-T5 and Raw T5's Rationale Following Abilities and the Effect of Different Types of Instructions

We separately trained raw T5 and instruction-tuned FLAN-T5 models using CoT and Relevant-Cell-based rationales, and compare the improvements over their respective baselines, as shown in Table 7. The results demonstrated that the results achieved by FLAN-T5 models through rationale-based training was comparable or better compared to raw T5 models, aligning with (Fu et al., 2023)'s observations. This suggests that, in future works that aim to enhance model reasoning and instruction-following abilities, utilizing instruction fine-tuned checkpoints of existing models could yield better performance.

We also examined the impact of different types of instructions, including "No Instructions", "Task Prefix instructions", and a different number of "Natural Language Instructions", as shown in Table 8. Results demonstrate that models trained without instructions achieved low performance of 28.1 sacreBLEU score, while those trained with 10 Natural Language instructions performed similarly to models trained with Task Prefix instructions, albeit slightly lower. Our Natural Language instructions were diverse and accurate descriptions of the two tasks in the multi-task learning framework, generated by the Teacher model and edited by the authors. The superior performance of the Task Prefix instruction might be attributed to T5's pre-training strategies, which employs similar task prefixes during pre-training on various tasks, such as `[summarize]`, `[snli sentence]`, `[cola sentence]`. This familiarity helps FLAN-T5 models better understand and distinguish the two tasks in our multi-task learning framework.

## 5 Conclusion

In conclusion, this paper introduces a novel Relevant-Cell-based Knowledge Distillation with inference-time Teacher Guidance method to combine small models' capabilities of learning from human annotation and LLMs' capabilities of reasoning effectively. Through a multi-task learning on instruction-tuned models and inferencing with Teacher model's guidance, we achieved a state-of-the-art BLEU score of 41.3 on the FeTaQA benchmark. This approach combines LLMs' robust reasoning strength with smaller models' task-specific capabilities, while maintaining high data and computational efficiency. Our work advances free-form table question answering and demonstrates a promising path of fusing general LLMs and task-specific models on downstream tasks. Future works could explore on how to more effectively combine and communicate between the Student model and Teacher model. Future works could also explore on extending our method to diverse downstream tasks, leveraging general LLMs as strong reasoners, and task-specific small models as response generators aligned with human annotated preferences.

CCL 2024

## References

Chandra Bhagavatula, Jena D. Hwang, Doug Downey, Ronan Le Bras, Ximing Lu, Keisuke Sakaguchi, Swabha Swayamdipta, Peter West, and Yejin Choi. 2022. I2d2: Inductive knowledge distillation with neurologic and self-imitation. In *Annual Meeting of the Association for Computational Linguistics*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, SHIYANG LI, Xiyou Zhou, and William Yang Wang. 2019. Tabfact: A large-scale dataset for table-based fact verification. *ArXiv*, abs/1909.02164.

Wenhu Chen, Jianshu Chen, Yunde Su, Zhiyu Chen, and William Yang Wang. 2020. Logical natural language generation from open-domain tables. *ArXiv*, abs/2004.10404.

Zhoujun Cheng, Haoyu Dong, Zhiruo Wang, Ran Jia, Jiaqi Guo, Yan Gao, Shi Han, Jian-Guang Lou, and Dongmei Zhang. 2021. Hitab: A hierarchical table dataset for question answering and natural language generation. *ArXiv*, abs/2108.06712.

Hyung Won Chung, Le Hou, S. Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Wei Yu, Vincent Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed Huai hsin Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *ArXiv*, abs/2210.11416.

Julian Eisenschlos, Syrine Krichene, and Thomas Müller. 2020. Understanding tables with intermediate pre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online. Association for Computational Linguistics.

Yao Fu, Hao-Chun Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. 2023. Specializing smaller language models towards multi-step reasoning. *ArXiv*, abs/2301.12726.

Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Martin Eisenschlos. 2020. Tapas: Weakly supervised table parsing via pre-training. In *Annual Meeting of the Association for Computational Linguistics*.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2022. Large language models are reasoning teachers. In *Annual Meeting of the Association for Computational Linguistics*.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander J. Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Annual Meeting of the Association for Computational Linguistics*.

Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2022. Large language models can self-improve. *ArXiv*, abs/2210.11610.

Mohit Iyyer, Wen tau Yih, and Ming-Wei Chang. 2017. Search-based neural structured learning for sequential question answering. In *Annual Meeting of the Association for Computational Linguistics*.

Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. In *Conference on Empirical Methods in Natural Language Processing*.

SHIYANG LI, Jianshu Chen, Yelong Shen, Zhiyu Chen, Xinlu Zhang, Zekun Li, Hong Wang, Jingu Qian, Baolin Peng, Yi Mao, Wenhu Chen, and Xifeng Yan. 2022. Explanations from large language models make small reasoners better. *ArXiv*, abs/2210.06726.


(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Liunian Harold Li, Jack Hessel, Youngjae Yu, Xiang Ren, Kai-Wei Chang, and Yejin Choi. 2023. Symbolic chain-of-thought distillation: Small models can also "think" step-by-step. In *Annual Meeting of the Association for Computational Linguistics*.

Qian Liu, Bei Chen, Jiaqi Guo, Zeqi Lin, and Jian-Guang Lou. 2021. Tapex: Table pre-training via learning a neural sql executor. *ArXiv*, abs/2107.07653.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022a. WANLI: Worker and AI collaboration for natural language inference dataset creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ao Liu, Haoyu Dong, Naoaki Okazaki, Shi Han, and Dongmei Zhang. 2022b. Plog: Table-to-logic pretraining for logical table-to-text generation. In *Conference on Empirical Methods in Natural Language Processing*.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2022. Teaching small language models to reason. In *Annual Meeting of the Association for Computational Linguistics*.

Nafise Sadat Moosavi, Andreas Rücklé, Dan Roth, and Iryna Gurevych. 2021. Scigen: a dataset for reasoning-aware text generation from scientific tables. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Linyong Nan, Chia-Hsuan Hsieh, Ziming Mao, Xi Victoria Lin, Neha Verma, Rui Zhang, Wojciech Kryscinski, Nick Schoelkopf, Riley Kong, Xiangru Tang, Murori Mutuma, Benjamin Rosand, Isabel Trindade, Renusree Bandaru, Jacob Cunningham, Caiming Xiong, and Dragomir R. Radev. 2021. Fetaqa: Free-form table question answering. *Transactions of the Association for Computational Linguistics*, 10:35–49.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155.

Ankur P. Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. Totto: A controlled table-to-text generation dataset. *ArXiv*, abs/2004.14373.

Panupong Pasupat and Percy Liang. 2015. Compositional semantic parsing on semi-structured tables. In *Annual Meeting of the Association for Computational Linguistics*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473.

Peng Shi, Patrick Ng, Feng Nan, Henghui Zhu, J. Wang, Jiarong Jiang, Alexander Hanbo Li, Rishav Chakravarti, Donald Weidner, Bing Xiang, and Zhiguo Wang. 2022. Generation-focused table-based intermediate pre-training for free-form question answering. In *AAAI Conference on Artificial Intelligence*.

Lya Hulliyyatus Suadaa, Hidetaka Kamigaito, Kotaro Funakoshi, Manabu Okumura, and Hiroya Takamura. 2021. Towards table-to-text generation with numerical reasoning. In *Annual Meeting of the Association for Computational Linguistics*.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications.

PeiFeng Wang, Aaron Chan, Filip Ilievski, Muhao Chen, and Xiang Ren. 2022. Pinto: Faithful language reasoning using prompt-generated rationales. In *The Eleventh International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Huai hsin Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903.


(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Peter West, Chandrasekhar Bhagavatula, Jack Hessel, Jena D. Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2021. Symbolic knowledge distillation: from general language models to commonsense models. In *North American Chapter of the Association for Computational Linguistics*.

Tianbao Xie, Chen Henry Wu, Peng Shi, Ruiqi Zhong, Torsten Scholak, Michihiro Yasunaga, Chien-Sheng Wu, Ming Zhong, Pengcheng Yin, Sida I. Wang, Victor Zhong, Bailin Wang, Chengzu Li, Connor Boyle, Ansong Ni, Ziyu Yao, Dragomir R. Radev, Caiming Xiong, Lingpeng Kong, Rui Zhang, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2022. Unifiedskg: Unifying and multi-tasking structured knowledge grounding with text-to-text language models. *ArXiv*, abs/2201.05966.

Jingfeng Yang, Aditya Gupta, Shyam Upadhyay, Luheng He, Rahul Goel, and Shachi Paul. 2022. Tableformer: Robust transformer modeling for table-text encoding. In *Annual Meeting of the Association for Computational Linguistics*.

Pengcheng Yin, Graham Neubig, Wen tau Yih, and Sebastian Riedel. 2020. Tabert: Pretraining for joint understanding of textual and tabular data. *ArXiv*, abs/2005.08314.

Tao Yu, Rui Zhang, Kai-Chou Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Z Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir R. Radev. 2018. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task. *ArXiv*, abs/1809.08887.

Yilun Zhao, Linyong Nan, Zhenting Qi, Rui Zhang, and Dragomir R. Radev. 2022. Reastap: Injecting table reasoning skills during pre-training via synthetic reasoning examples. In *Conference on Empirical Methods in Natural Language Processing*.

Yilun Zhao, Boyu Mi, Zhenting Qi, Linyong Nan, Minghao Guo, Arman Cohan, and Dragomir R. Radev. 2023a. Openrt: An open-source framework for reasoning over tabular data. In *Annual Meeting of the Association for Computational Linguistics*.

Yilun Zhao, Zhenting Qi, Linyong Nan, Lorenzo Jaime Yu Flores, and Dragomir R. Radev. 2023b. Loft: Enhancing faithfulness and diversity for table-to-text generation via logic form control. In *Conference of the European Chapter of the Association for Computational Linguistics*.

Yilun Zhao, Zhenting Qi, Linyong Nan, Boyu Mi, Yixin Liu, Weijin Zou, Simeng Han, Xiangru Tang, Yumo Xu, Arman Cohan, and Dragomir R. Radev. 2023c. Qtsumm: A new benchmark for query-focused table summarization. *ArXiv*, abs/2305.14303.

Victor Zhong, Caiming Xiong, and Richard Socher. 2017. Seq2sql: Generating structured queries from natural language using reinforcement learning. *CoRR*, abs/1709.00103.

CCL 2024

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 973-985, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

# Enhancing Sequence Representation for Personalized Search

**Shijun Wang**
School of Information
Renmin University of China
wangshijun@ruc.edu.cn

**Han Zhang**
China Unicom Research Institute
zhangh1550@chinaunicom.cn

**Zhe Yuan**
Academy of Regional and Global Governance
Beijing Foreign Studies University
yuanzhe@bfsu.edu.cn

## Abstract

The critical process of personalized search is to reorder candidate documents of the current query based on the user's historical behavior sequence. There are many types of information contained in user historical information sequence, such as queries, documents, and clicks. Most existing personalized search approaches concatenate these types of information to get an overall user representation, but they ignore the associations among them. We believe the associations of different information mentioned above are significant to personalized search. Based on a hierarchical transformer as base architecture, we design three auxiliary tasks to capture the associations of different information in user behavior sequence. Under the guidance of mutual information, we adjust the training loss, enabling our PSMIM model to better enhance the information representation in personalized search. Experimental results demonstrate that our proposed method outperforms some personalized search methods.

## 1 Introduction

Search engines have emerged as pivotal instruments for users to access information. However, search engines often yield identical results in response to identical queries, regardless of different information needs from different users, e.g., some users use the word "MAC" to refer to "MAC computer", in contrast, some other users use it to denote "MAC lipstick". To tackle this problem, personalized search has emerged as a highly effective strategy. It models a user's preferences by user's historical behavior, and returns customized ranking results. Personalized search has been proved to provide a better re-ranking document list and improve user satisfaction (Dou et al., 2007; Bennett et al., 2012; White et al., 2013; Cai et al., 2014; Ge et al., 2018; Ma et al., 2020; Zhou et al., 2020b).

Prevalent models integrate semantic context from user histories with current queries and documents. Deep learning models like hierarchical recurrent network with query-aware attention and graph-augmented approach have advanced the field but typically neglect the interconnectedness of historical informational units (Ge et al., 2018; Lu et al., 2020).

We believe that capturing contextual relationships between historical queries/documents and their behavioral contexts is vital for better personalization. Hence, we introduce an innovative methodology rooted in Mutual Information Maximization (MIM, elaborated upon in Appendix A). As a form of self-supervised learning, MIM exploits the inherent correlations present in user historical sequences to blend different types of data into a cohesive representation. This learning paradigm (Devlin et al., 2019; Zhou et al., 2020a) enables models to learn from the underlying structure of raw data, creating self-derived training signals and initializing model parameters through specifically designed optimization objectives.

We present the PSMIM (Personalized Search Mutual Information Maximization) model, incorporating auxiliary tasks to enhance user representations for personalized search. By exploiting self-supervised information from user historical sequences, we design three auxiliary tasks to capture the following correlations: (1) the correlation between historical query and historical sequence information, (2) the correlation between historical document and historical sequence information, and (3) the correlation between

©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 986-998, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

the current query and historical sequence information. These objectives are unified under the framework of maximizing mutual information, enabling the model to uncover associations among different information types in user sequences and adapt to new data or association patterns. By effectively fusing various user information types through auxiliary tasks, we obtain a richer user historical context representation, which is then integrated with candidate document representations.

We employ Transformers to capture interactions among the historical behavior sequence, current query, and candidate documents to compute personalized ranking scores. Our model is trained with pair-wise ranking loss. Through experiments, we validate the effectiveness of our approach, demonstrating that our model, PSMIM, outperforms several previous personalized search models.

Our contributions include proposing a personalized search model PSMIM which systematically addresses inter-associations among user sequence information through auxiliary objectives, and empirically validating the practical effectiveness and robustness of PSMIM.

## 2 Related Work

### 2.1 Personalized Web Search

The essence of personalized search lies in deciphering users' preferences and traits based on their historical interactions. Hence, scholars have developed several techniques to extract user-specific details from historical sequences. Initially, research efforts such as (Dou et al., 2007; Teevan et al., 2011) emphasized click behavior in user history, proving it to be a straightforward yet insightful tool for predicting user inclinations. Moreover, many studies (Sieg et al., 2007; White et al., 2013; Harvey et al., 2013; Vu et al., 2015; Vu et al., 2017) endeavored to formulate user profiles by examining the thematic attributes of documents within the user's search history. However, they faced challenges, including high computational expenses and incomplete categorization.

Recently, deep learning has garnered broad acceptance across disciplines due to its exceptional representation of learning capabilities. It successfully uncovers nuanced user preferences in personalized search contexts, as documented in (Ge et al., 2018; Lu et al., 2019; Yao et al., 2020b; Lu et al., 2020; Yao et al., 2020a; Ma et al., 2020; Zhou et al., 2020b).

Distinctively, our methodology capitalizes on the interdependencies among historical queries, historical documents, the current query, and the overall user historical sequence. By doing so, we aim to refine their representations and thoroughly explore the dynamics between candidate documents and the user's historical context, thus contributing to advancements in personalized search.

### 2.2 Self-supervised Learning

Self-supervised learning techniques (Devlin et al., 2019; Hjelm et al., 2019) leverage auxiliary tasks to extract meaningful insights from unsupervised data. By generating supervised signals from intrinsic data relationships, these methods foster effective representations for subsequent tasks. Some self-supervised frameworks (Hjelm et al., 2019) utilize naturally associated features for visual feature learning. Pre-trained language models (PLMs) (Devlin et al., 2019) represent a prominent example of self-supervised applications in the NLP domain.

Mutual Information Maximization (MIM) (Hjelm et al., 2019; Kong et al., 2020) is a self-supervised learning technique. It operates by splitting input data into multiple perspectives and strives to maximize mutual information among the representations of these perspectives.

Unlike prior strategies, our work considers correlations among user search history and contextual information as a self-supervised cue. We present a model utilizing MIM to enhance sequence-level user information embeddings, thereby boosting personalized search task performance.

## 3 Problem Definition

First, let us define the personalized search problem. Consider a user's historical information sequence $H = \{P_1, P_2, \dots\}$, with each page $P_i$ contains an issued query, a set of documents, and a clicked document with its timestamp. Acknowledging that issued queries and their corresponding clicked documents are more informational, so our study focuses primarily on modeling these two critical types of historical

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 986-998, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

data within the sequence. Thus, we can assume a user's history as $H = \{(q_1, d_1), \ldots, (q_n, d_n)\}$, with $n$ representing the total count of queries in the user's history. Given a current query $q_c$ and a set of candidate documents $D_c = \{d_{c_1}, d_{c_2}, \ldots, d_{c_m}\}$, the personalized search task is to score each document in $D_c$ according to the current query $q_c$ and the user's historical information sequence $H$.

## 4 Our Method

Our method maximizes mutual information across user historical data, including historical queries, historical documents, alongside the current query. Auxiliary tasks help to enhance the model's capacity to represent diverse historical information types. Leveraging word-level and sequence-level transformers, our method effectively encodes the contextual and sequential nuances within user historical data.

### 4.1 Base Model

Here we introduce the base model upon which our contributions are built. It serves as a solid foundation to present our method. Due to word polysemy and contextual variations, user historical informaiton reflects diverse user intentions and interests. So we design it as a hierarchical transformer architecture, illustrated in Figure 1, which is crucial address the complexities of user historical information.

User historical information includes queries, documents, and timestamps. We use a word-level transformer to fuse the context information of the words in a query or a document, and use a sequence-level transformer to fuse the context information of the queries and the documents, then get a context-aware representation for a user information sequence. This architecture considers context at both the word and sequence levels, allowing us to capture the nuances of user interactions with the system over time.

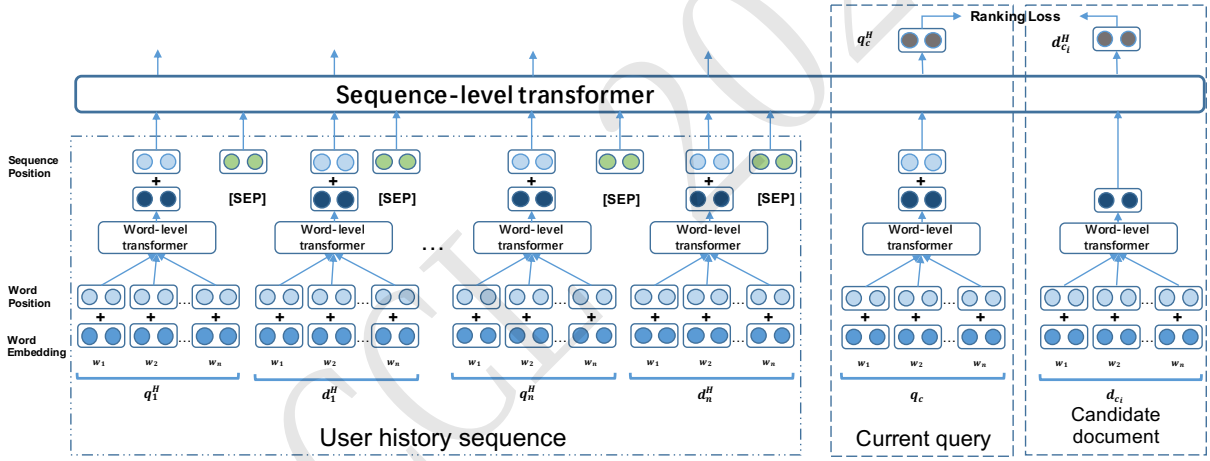



Figure 1: Base model of PSMIM (Personalized Search Mutual Information Maximization) framework.

The main component of the base model is a two-level context encoder, serving as context encoder for user history sequence.

**The word-level context encoder** is a transformer that takes in words from a query or a clicked document, and encodes them into a query embedding or a document embedding, preparing for the sequence-level context encoder. The goal of the word-level context encoder is to learn better representation for a query or a document, based on its whole text.

Taking a query $q$ as an example (the word-level representation of document is the same as that of a query): suppose it consists of $m$ words, expressed as $q = \{w_1, w_2, \ldots, w_m\}$. The word-level transformer takes in the word list $\{w_1, w_2, \ldots, w_m\}$, and then we get the embedding of query $q$.

**The sequence-level context encoder** is a transformer that takes in word-level embeddings of queries and documents from user history sequence, and encodes them into a user historical information sequence embedding. The goal of the sequence-level context encoder is to learn better representation for a user's historical information sequence, in other words, user history.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 986-998, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

As we discussed earlier, the queries are usually very concise, sometimes just one word. In this case, there will still be ambiguity. Users usually put forward a series of queries for a single information requirement. In this process, historical queries and clicked documents provide rich context information to infer the user's intention. Therefore, we consider introducing the historical sequence context information to help represent the user's queries. Similarly, historically clicked documents represent the user's interest, and the combination of context query and document can better represent the user's interest in the documents.

After the word-level context encoder, a query or a document is integrated with the word-level context information. We take all queries and documents as context information to get the sequence-level representation. Suppose that the user historical information sequence is represented as the following list:

$$H = \{q_1, d_1, \ldots, q_t, d_t, \ldots, q_n, d_n, q_c\}, \tag{1}$$

the sequence-level transformer takes in the sequence list, then we get the embedding of the user historical information sequence $H$.

**Current query representation**: After the current query is sent into the Word-level Transformer as shown in Figure 1, each word in the current query is integrated with the word context information in the query. We conduct sum-pooling for all word representations of the query as the final representation. After the Sequence-level Transformer, the current query representation $q_c^H$ incorporates information about the user's historical intentions and interests from the historical sequence. It should be noted that when the current query is fed into the Sequence-level Transformer, we only feed the historical information sequence into the model, and we do not feed the candidate document into the Transformer model.

**Candidate document representation**: Similar to the processing of the current query, after the candidate document is processed through the Word-level Transformer, we get a document representation that incorporates the word context, and after the sequence-level Transformer, the candidate document representation $d_{c_i}^H$ is represented with information that incorporates the user's historical intentions and interests. Similarly, when candidate documents are sent into the Sequence-level Transformer, we only consider historical information sequences in the model and do not use the current query in the Transformer model.

### 4.2 Auxiliary Learning Tasks

On the basis above, we design three auxiliary tasks with mutual information maximization using self-supervised information from user historical sequence to enhance the representation of user information. These three auxiliary tasks are respectively historical query prediction, historical document prediction and current query prediction. They model the associations of different information among the user historical sequence with mutual information maximization. These three auxiliary tasks are constructed as shown in Figure 2.

#### 4.2.1 Historical Query Prediction

A historical query provides the user's fine-grained intention at a specific time. Our goal is establishing the correlation between a historical query and the user's whole historical sequence, to better represent the historical query. Existing personalized search models usually concatenate historical queries and documents, then regrade user historical information as a from-left-to-right sequence, which ignores the associations between historical queries and the whole sequence. It is also noted that during the training process, the model can observe the entire sequence of user historical information. Inspired by the model of BERT, we propose to use the bidirectional information in the historical sequence by a cloze task.

As an illustration, given a user historical sequence $H$ and a historical query $q_t$, we mask the t-th query $q_t$. Then we take the rest sequence as its surrounding context, which is represented as:

$$C_{q_t} = \{q_1, d_1, \ldots, [mask], d_t, \ldots, q_n, d_n, q_c\}. \tag{2}$$

Considering the surrounding context $C_{q_t}$ and masking term $q_t$, we treat them as distinct views for learning the data representation. According to the equations 19 in Appendix A, we minimize the Historical Query

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 986-998, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

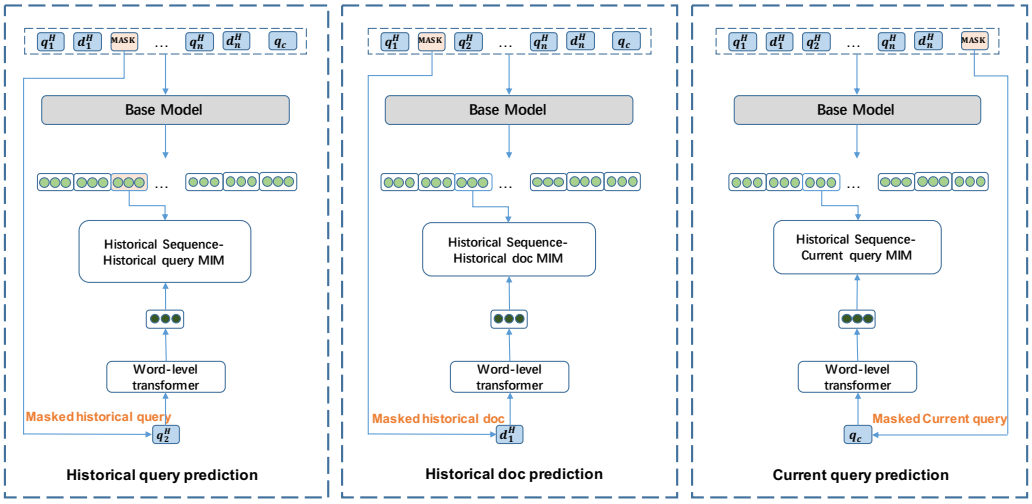



Figure 2: Three auxiliary tasks (Historical query prediction, Historical document prediction and Current query prediction) to model the associations of different information among the user historical sequence with Mutual Information Maximization to enhance the representation of user information.

Prediction (HQP) loss by:

$$L_{\mathrm{HQP}}(C_{q_t}, q_t) = f(C_{q_t}, q_t) - \log\left(\sum_{\tilde{q}\in\mathcal{Q}\setminus\{q_t\}} \exp f(C_{q_t}, q_t)\right), \tag{3}$$

where $\tilde{q}$ denotes the negative sampling query and $f(\cdot)$ is calculated according to the following formula:

$$f(C_{q_t}, q_t) = \sigma(F^S_{q_t} \cdot W_{\mathrm{HQP}} \cdot q^W_t), \tag{4}$$

where $W_{\mathrm{HQP}} \in \mathbb{R}^*$ is the parameter matrix to be learned, $F^S_{q_t}$ is the learned sequence-level representation using the t-th query position of the user historical information sequence by sequence-level transformer, and $q^W_t$ is the sequence-level representation of query $q_t$.

### 4.2.2 Historical Document Prediction

A clicked documents offers some fine-grained insights matching to the user's past interest in at a specific time. Our goal is establishing the correlation between a historical clicked document and the user's whole historical sequence, to better represent the clicked document.

Consistently with the previous subsection, given a user historical sequence $H$, we mask the t-th clicked document $d_t$, and take the rest sequence as its surrounding, and it is expressed as:

$$C_{d_t} = \{q_1, d_1, \ldots, q_t, [mask], \ldots, q_n, d_n, q_c\}. \tag{5}$$

We combine the context $C_{d_t}$ and masking $d_t$ as two different views. According to the equations in the 19 in Appendix A, we minimize the historical document prediction (HDP) loss by:

$$L_{\mathrm{HDP}}(C_{d_t}, d_t) = f(C_{d_t}, d_t) - \log\left(\sum_{\tilde{d}\in\mathcal{D}\setminus\{d_t\}} \exp f(C_{d_t}, d_t)\right), \tag{6}$$

where $\tilde{d}$ denotes the negative sampling document and $f(\cdot)$ is calculated by:

$$f(C_{d_t}, d_t) = \sigma(F^S_{d_t} \cdot W_{\mathrm{HDP}} \cdot d^W_t), \tag{7}$$

where $W_{\mathrm{HDP}} \in \mathbb{R}^{d*d}$ is the parameter matrix to be learned, $F^S_{d_t}$ is the learned sequence-level representation using the t-th document position of the user historical information sequence by sequence-level transformer, and $d^W_t$ is the word-level representation of document $d_t$.

CCL 2024

#### 4.2.3 Current Query Prediction

The current query reflects the user's current intent. Our goal is establishing the correlation between the current query and the user's whole historical sequence, to better represent the user's current intent. Similarly, we mask the current query $q_c$ and take the rest of the information sequence as its surrounding context $C_{q_c}$, then we minimize the current query prediction (CQP) loss by:

$$L_{\mathrm{CQP}}(C_{q_c}, q_c) = f(C_{q_c}, q_c) - \log\left(\sum_{\tilde{q}\in\mathcal{Q}_{\mathcal{H}}} \exp f(C_{q_c}, q_c)\right), \tag{8}$$

where $\tilde{q}$ denotes the negative query sample and $f(\cdot)$ is calculated according to the following formula:

$$f(C_{q_c}, q_c) = \sigma(q_c^S \cdot W_{\mathrm{CQP}} \cdot q_c^W), \tag{9}$$

where $W_{\mathrm{CQP}} \in \mathbb{R}^{d*d}$ is the parameter matrix to be learned, $F_{q_c}^S$ is the learned sequence-level representation at the current query position in the user information sequence by sequence-level transformer, and $q_c^W$ is the word-level representation of query $q_c$.

### 4.3 Candidate Document Score

In our method, given a current query $q$ and a candidate document $d$, the score of the candidate document consists of two parts: the personalized relevance $p(d^H, q^H)$ and the ad-hoc relevance $p(d, q)$. As for the personalized relevance $p(d^H, q^H)$, we obtain the $d^H$ of a candidate document integrating historical sequence information and the $q^H$ of the current query through our model. After sum-pooling, our model acquire the integrated word-level context representation $d^{W_{sum}}$, upon which we conduct interaction with the current query $q^H$. Finally, we use one hidden layer MLP to automatically adjust the weights of various parts of these scores as below:

$$p(d^H, q^H) = \Phi(S(q^H, d^H), S(q^H, d^{W_{sum}})), \tag{10}$$

where, $S(\cdot)$ calculates similarity based on vector representation. In our work, we specifically employ cosine similarity.

The ad-hoc relevance $p(d, q)$ is computed through term matching between the current query and the candidate document, augmented by the context-aware word-level representations $q^W$ and $d^W$. In addition, we extract the manual feature $F(q, d)$ of the clicks and topic for each document. These features are also weighted using one hidden layer MLP, as shown below:

$$p(d, q) = \Phi(I(q, d), I(q^W, d^W), \phi(F(q, d))), \tag{11}$$

where $I(\cdot)$ is based on the similarity of interactions. We adopt the KRNM (Xiong et al., 2017) method to implement the interaction. Specifically, given a query-document pair, we use the cosine similarity of each term in the query and document to construct an interaction matrix $M$. Then we use $k$ kernels for matches. The final matching score $\mathrm{KNRM(M)}$ is obtained by aggregating them using one hidden layer MLP.

$$\mathrm{KNRM_k(M)} = \phi(\mathrm{K_1(M)}, \mathrm{K_2(M)}, \ldots, \mathrm{K_k(M)}), \tag{12}$$

$$K_k(M) = \sum_i \log\left(\sum_j \exp\left(\frac{(M_{ij} - \mu_k)^2}{2\sigma_k^2}\right)\right), \tag{13}$$

where $\mu_k$ is distributed between -1 and 1 depending on the number of kernels, and $\sigma_k$ is set to 0.1.

Finally, we re-rank the candidate documents according to the final relevance score to obtain personalized search results that meet the users' needs.

CCL 2024

### 4.4 Optimization Objective

In this section, we detail our model's optimization objective and the parameter configurations.

**Ranking Loss**. Many algorithms train the ranking task in pair-wise mode (Bennett et al., 2012; White et al., 2013; Ge et al., 2018; Lu et al., 2019; Yao et al., 2020b; Lu et al., 2020; Yao et al., 2020a; Ma et al., 2020; Zhou et al., 2020b). Each pair consists of a clicked document and an unclicked document, and the optimization goal is to maximize the gap in their scores. In particular, given a clicked document $d_i$ and an unclicked document $d_j$, $d_i$ is more relevant than $d_j$, and the prediction probability is $p(d_i|q, H) - p(d_j|q, H)$. The Ranking loss is the weighted cross-entropy between the ground-truth label $\overline{p}_{ij}$ and the predicted probability $p_{ij}$:

$$L_{\text{rank}} = -|\lambda_{ij}| \left(\overline{p}_{ij} \log p_{ij} + \overline{p}_{ji} \log p_{ji}\right), \tag{14}$$

where the weighting factor $\lambda_{ij}$ represents the impact on ranking quality when documents $d_i$ and $d_j$ swap positions.

**Auxiliary Task Losses**. As mentioned above, we designed three self-supervised auxiliary tasks, each associated with a loss function: $L_{\text{HQP}}$, $L_{\text{HDP}}$, and $L_{\text{CQP}}$. We minimize the loss functions of the three auxiliary tasks together to update the parameters of our model. It should be noted that our model parameters are shared for the auxiliary and ranking tasks. The total loss of auxiliary tasks is as follows:

$$L_{\text{aux}} = z_1 L_{\text{HQP}} + z_2 L_{\text{HDP}} + z_3 L_{\text{CQP}}, \tag{15}$$

where $z_1$, $z_2$ and $z_3$ are the hyperparameters that control the weights.

Finally, we minimize the combination of these two losses, adjusted by a hyper-parameter $\alpha$:

$$L_{\text{total}} = L_{\text{rank}} + \alpha L_{\text{aux}}. \tag{16}$$

## 5 Experiments

### 5.1 Dataset and Evaluation Metrics

Our study conducts experiments on the AOL dataset (Pass et al., 2006). The statistical details of the AOL dataset are summarized in Table 1.

Table 1: Statistics of the dataset

| Type | AOL dataset |
|---|---|
| #Days | 91 |
| #Users | 110,439 |
| #Queries | 736,454 |
| Average query length | 2.87 |
| Average #click per query | 1.11 |

The AOL dataset includes three months of user queries and corresponding clicked documents, fitting for personalized search purposes. Each user record includes an anonymous user ID, session ID, query, document, and click labels. We partition each user's data temporally into historical and experimental datasets. Historical data, encompassing the initial five weeks, serves as personalized historical information for each user's corresponding experimental data. The experimental data, covering the final eight weeks, is subsequently divided into training, validation, and testing sets following a 6:1:1 split ratio.

Given that the AOL dataset provides only clicked documents, we adhere to the methodology outlined in (Ahmad et al., 2018) and utilize the BM25 algorithm to identify the top-scoring candidate documents. Similarly, adopting the practices in (Ahmad et al., 2019; Huang et al., 2018), we extract 50 candidate documents per query in the testing set and five candidates per query in the training and validation sets. To ensure compliance with the fundamental principles of personalized search, we eliminate users whose historical data or training set is void. Furthermore, we exclusively use document titles for relevance calculations. In our study, we classify clicked documents in the AOL dataset as relevant, while the rest irrelevant. To assess the performance of various models, we employ three widely used metrics that gauge ranking quality: Mean Average Precision (MAP), Mean Reciprocal Rank (MRR), and Precision@1 (P@1). More implementation details are listed in Appendix B.

CCL 2024

### 5.2 Baselines

We display the original ranking performance from the dataset and take several traditional and personalized search models as baseline models.

**BM25** (Robertson and Zaragoza, 2009). As mentioned above, we use the BM25 algorithm to get one fundamental ranking.

**KNRM** (Xiong et al., 2017). This model conducts a kernel-pooling technique on a word-level similarity matrix to extract soft match features in multiple granularities. These features are then exploited by a pairwise Learning-to-Rank (LTR) algorithm to derive the final ranking score.

**Conv-KNRM** (Dai et al., 2018). In this model, a convolution layer is added based on KNRM to construct the n-gram soft matches, and the context information of words is added.

**Bert** (Devlin et al., 2019). We input the query and document simultaneously into a pre-trained BERT model and extract the final "[CLS]" token's output as the matching feature to compute the matching score.

**P-Click** (Dou et al., 2007). This model utilizes the click frequency and initial position of candidate documents in the user's historical sequence as personalized features for document re-ranking It has a good performance on user's refinding behavior.

**SLTB** (Bennett et al., 2012). In this model, click features, topic features, time features, and location features are taken as the features of personalized search. It utilizes a learning-to-rank approach to achieve document re-ranking.

**HRNN** (Ge et al., 2018). The model uses the hierarchical recurrent neural network and attention mechanism to model the historical information related to the current query and constructs the user profile based on the current query for the personalized search.

**PSTIE** (Ma et al., 2020). This model considers using time interval information in history to enhance the performance of the personalized search.

**HTPS** (Zhou et al., 2020b). This model divides the user history into long-term and short-term histories to disambiguate the current query and designs a personalized language model task to improve the model's effectiveness. Since the word embedding of most personalized search methods is fixed, we keep the word embedding static without fine-tuning for a fair comparison.

### 5.3 Experimental Results

We categorize our baseline models into two main groups: traditional search methods and personalized search methods. As shown in Table 2, our proposed personalized search model PSMIM surpasses traditional search models and other personalized search methodologies, outperforming several counterparts.

Table 2: Overall performance of all models on AOL dataset. "†" indicates the model outperforms all baselines significantly with paired t-test at $p < 0.05$ level. Best results are shown in bold.

| Task | Model | MAP | | MRR | | P@1 | |
|---|---|---|---|---|---|---|---|
| Traditional Search | BM25 | .250 | -62.80% | .258 | -62.40% | .148 | -75.05% |
| | KNRM | .429 | -36.16% | .439 | -36.01% | .271 | -54.31% |
| | Conv-KNRM | .474 | -29.46% | .485 | -29.31% | .327 | -44.86% |
| | BERT | .483 | -28.13% | .493 | -28.14% | .335 | -43.51% |
| Personalized Search | P-Click | .422 | -37.20% | .431 | -37.17% | .378 | -36.26% |
| | SLTB | .511 | -23.96% | .524 | -23.62% | .469 | -20.92% |
| | HRNN | .544 | -19.05% | .557 | -18.80% | .484 | -18.39% |
| | PSTIE | .564 | -16.07% | .577 | -15.89% | .503 | -15.18% |
| | HTPS | .672 | - | .686 | - | .593 | - |
| | PSMIM | **.689**$^{\dagger}$ | +2.53% | **.703**$^{\dagger}$ | +2.48% | **.610**$^{\dagger}$ | +2.87% |

Subsequently, these personalized search methods can be divided into two categories: those based on user profiles (P-Click, SLTB, HRNN and PSTIE), and context-aware approaches (HTPS and PSMIM). By comparing context-aware personalized methods with those based on user profiles, we find that HTPS and PSMIM are superior to previous personalized search models. This finding emphasizes the importance of historical context.

CCL 2024

Further, we contrast our model PSMIM against context-aware model HTPS. As evidenced in Table 2, our model excels over preceding personalized search models. Compared with the benchmark HTPS, our model exhibits statistically significant enhancements across all evaluation metrics, as confirmed by paired t-test at a $p < 0.05$ significance level. Significantly, on the AOL dataset, PSMIM records a notable improvement over HTPS with a 2.53% increase in MAP, a 2.48% boost in MRR, and a 2.87% rise in P@1. These results prove that our method is effective and robust.

In conclusion, the experimental outcomes affirm that accounting for the interconnections within user sequence information and employing context-aware candidate document representations effectively enhances the performance of personalized search. To further dissect the functionality of our model components, we conduct an ablation study and some further experiments in the upcoming subsections.

## 5.4 Discussion

### 5.4.1 Ablation Study

Our PSMIM model includes several essential components: the current query prediction task (CQP), the historical query prediction task (HQP), the historical document prediction task (HDP), and candidate document representation as historical context information (CDC). To prove the effectiveness of these main components in the PSMIM model, we perform several ablation experiments. We present the experimental results in Table 3 and have some discussions. Specifically, in setting PSMIM without CQP,

Table 3: Results of ablation experiments.

| Model | MAP | MRR | P@1 |
|---|---|---|---|
| PSMIM | **0.6887** | **0.7029** | **0.6096** |
| w/o. CQP | 0.6789 | 0.6932 | 0.5976 |
| w/o. HQP&HDP | 0.6713 | 0.6853 | 0.5928 |
| w/o. CDC | 0.6800 | 0.6939 | 0.5999 |

we remove the current query prediction task of the PSMIM model and then train and test the model. In setting PSMIM without HQP&HDP, since the historical query and historical document are inseparable sources for deriving context information, we concurrently remove the historical query prediction and historical document prediction tasks from the original process. In setting PSMIM without CDC, we replace the candidate document representation integrating historical context information with word embedding representation of the original candidate document.

In Table 3, it is evident that the performance of our model diminishes under the three implemented ablation strategies, falling short of the complete model's performance. Notably, the removal of historical query and historical document prediction tasks leads to the most substantial decrease in MAP, confirming the necessity and significance of capturing associations within users' historical interaction sequences. Discontinuing the current query prediction task also impairs the results, pointing to the importance of strengthening the representation of the current query for improved performance. Moreover, excluding historical context into candidate document embeddings leads to reduced model performance, highlighting the benefits of exploiting personalized candidate document representations.

### 5.4.2 Effect of Short-term and Long-term History

To validate the robustness of our model, akin to HTPS (Zhou et al., 2020b), we partition user history into short-term and long-term segments, and scrutinize the performance of our model under both short-term and combined short-term and long-term histories scenarios.

**HTPS-S**. We select HTPS, a model surpassing most prior personalized search approaches, as our baseline. We only use short-term history during training and testing the model.

**PSMIM-S**. We only use short-term user history during training and testing our model.

The results are shown in Table 4. It could be seen from the experimental results that our model PSMIM with all user history significantly outperforms the model HTPS (Zhou et al., 2020b), and the model PSMIM-S with short-term history still outperforms the model HTPS in terms of MAP and MRR, which can prove the robustness of our model.

CCL 2024

Table 4: Performance of models with short and long term historys. † indicates the model outperforms baselines significantly with paired t-test at p-value $< 0.05$.

| Model | MAP | MRR | P@1 |
|---|---|---|---|
| HTPS-S | 0.6280 | 0.6410 | 0.5210 |
| HTPS | 0.6720 | 0.6860 | 0.5930 |
| PSMIM-S | 0.6737† | 0.6877† | 0.5921 |
| PSMIM | **0.6887**† | **0.7029**† | **0.6096**† |

However, PSMIM-S exhibits marginally lower performance than HTPS (Zhou et al., 2020b) on P@1 metric. A possible reason is that the model PSMIM-S with the short-term history could not model the overall historical interests of a user.

### 5.4.3 Performance of Ambiguous and Non-ambiguous Query Sets

Typically, a user's queries fall into two categories: navigational and informational. The navigational queries usually have more precise query intent, and the informational queries are usually more ambiguous. Studies [8] have shown that it is more necessary to personalize the search results for informational queries, which typically exhibit higher click entropy. To verify the robustness of our model, we follow the setting in [29], computing click entropy for queries, using a threshold of 1.0 to separate queries into two subsets, and subsequently contrasting the results with the baseline model HTPS.

From the results in Table 5, our model outperforms the baseline model HTPS, especially on the queries with larger click-entropy. It indicates that our model can learn better user information representation when facing ambiguous queries.

Table 5: Performance of Ambiguous and Non-Ambiguous Query Sets.

| Query sets | on Ambiguous Query Set | on Non-Ambiguous Query Set |
|---|---|---|
| HTPS | 0.3322 | 0.4240 |
| PSMIM | 0.3540 | 0.4421 |

### 5.4.4 Performance of Repeated and Non-repeated Query Sets

A user's queries could be classified into repeated and non-repeated queries. It is easy to infer the user's intent based on the same queries in the user history for repeated queries. Non-repeated queries, however, offer limited information based on their historical clicked documents. In this situation, a better query representation may bring some performance benefits.

As the results are shown in Table 6, our model outperforms the baseline model HTPS on both repeated queries and non-repeated queries. This indicates that our model can better infer the intent when facing a new query with a better information representation.

Table 6: Performance of Repeated and Non-Repeated Query sets.

| Model | on Repeated Query Sets | on Non-Repeated Query Sets |
|---|---|---|
| HTPS | 0.6048 | 0.2576 |
| PSMIM | 0.6310 | 0.2908 |

## 6 Conclusion

In this paper, we propose a personalized search model PSMIM under the guidance of mutual information, and design three auxiliary tasks to obtain the associations among different types of data from user history. Experiment results verify the effectiveness and robustness of our personalized search model PSMIM.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 986-998, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

## Acknowledgements

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve this paper. The study was supported by the China Postdoctoral Fellowship Program of CPSF(GZC20230287) and the Fundamental Research Funds for the Central Universities(2024QY004).

## References

Wasi Uddin Ahmad, Kai-Wei Chang, and Hongning Wang. 2018. Multi-task learning for document ranking and query suggestion. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Wasi Uddin Ahmad, Kai-Wei Chang, and Hongning Wang. 2019. Context attentive document ranking and query suggestion. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 385–394, New York, NY, USA. Association for Computing Machinery.

Paul N. Bennett, Ryen W. White, Wei Chu, Susan T. Dumais, Peter Bailey, Fedor Borisyuk, and Xiaoyuan Cui. 2012. Modeling the impact of short- and long-term behavior on search personalization. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '12, page 185–194, New York, NY, USA. Association for Computing Machinery.

Fei Cai, Shangsong Liang, and Maarten de Rijke. 2014. Personalized document re-ranking based on bayesian probabilistic matrix factorization. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, page 835–838, New York, NY, USA. Association for Computing Machinery.

Zhuyun Dai, Chenyan Xiong, Jamie Callan, and Zhiyuan Liu. 2018. Convolutional neural networks for soft-matching n-grams in ad-hoc search. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, WSDM '18, page 126–134, New York, NY, USA. Association for Computing Machinery.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Zhicheng Dou, Ruihua Song, and Ji-Rong Wen. 2007. A large-scale evaluation and analysis of personalized search strategies. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, page 581–590, New York, NY, USA. Association for Computing Machinery.

Songwei Ge, Zhicheng Dou, Zhengbao Jiang, Jian-Yun Nie, and Ji-Rong Wen. 2018. Personalizing search results using hierarchical rnn with query-aware attention. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, page 347–356, New York, NY, USA. Association for Computing Machinery.

Michael U. Gutmann and Aapo Hyvärinen. 2012. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *J. Mach. Learn. Res.*, 13(null):307–361, February.

Morgan Harvey, Fabio Crestani, and Mark J. Carman. 2013. Building user profiles from topic models for personalised search. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, CIKM '13, page 2309–2314, New York, NY, USA. Association for Computing Machinery.

R. Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Philip Bachman, Adam Trischler, and Yoshua Bengio. 2019. Learning deep representations by mutual information estimation and maximization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Jizhou Huang, Wei Zhang, Yaming Sun, Haifeng Wang, and Ting Liu. 2018. Improving entity recommendation with search log and multi-task learning. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4107–4114. International Joint Conferences on Artificial Intelligence Organization, 7.

Lingpeng Kong, Cyprien de Masson d'Autume, Lei Yu, Wang Ling, Zihang Dai, and Dani Yogatama. 2020. A mutual information maximization perspective of language representation learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Shuqi Lu, Zhicheng Dou, Xu Jun, Jian-Yun Nie, and Ji-Rong Wen. 2019. Psgan: A minimax game for personalized search with limited and noisy click data. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 555–564, New York, NY, USA. Association for Computing Machinery.

Shuqi Lu, Zhicheng Dou, Chenyan Xiong, Xiaojie Wang, and Ji-Rong Wen. 2020. Knowledge enhanced personalized search. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 709–718, New York, NY, USA. Association for Computing Machinery.

Zhengyi Ma, Zhicheng Dou, Guanyue Bian, and Ji-Rong Wen. 2020. Pstie: Time information enhanced personalized search. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 1075–1084, New York, NY, USA. Association for Computing Machinery.

Paninski and Liam. 2014. Estimation of entropy and mutual information. *Neural Computation*, 15(6):1191–1253.

Greg Pass, Abdur Chowdhury, and Cayley Torgeson. 2006. A picture of search. In *Proceedings of the 1st International Conference on Scalable Information Systems*, InfoScale '06, page 1–es, New York, NY, USA. Association for Computing Machinery.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389, apr.

Ahu Sieg, Bamshad Mobasher, and Robin Burke. 2007. Web search personalization with ontological user profiles. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, page 525–534, New York, NY, USA. Association for Computing Machinery.

Jaime Teevan, Daniel J. Liebling, and Gayathri Ravichandran Geetha. 2011. Understanding and predicting personal navigation. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, page 85–94, New York, NY, USA. Association for Computing Machinery.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748.

Thanh Vu, Alistair Willis, Son N. Tran, and Dawei Song. 2015. Temporal latent topic user profiles for search personalisation. In Allan Hanbury, Gabriella Kazai, Andreas Rauber, and Norbert Fuhr, editors, *Advances in Information Retrieval*, pages 605–616, Cham. Springer International Publishing.

Thanh Vu, Dat Quoc Nguyen, Mark Johnson, Dawei Song, and Alistair Willis. 2017. Search personalization with embeddings. In Joemon M Jose, Claudia Hauff, Ismail Sengor Altıngovde, Dawei Song, Dyaa Albakour, Stuart Watt, and John Tait, editors, *Advances in Information Retrieval*, pages 598–604, Cham. Springer International Publishing.

Ryen W. White, Wei Chu, Ahmed Hassan, Xiaodong He, Yang Song, and Hongning Wang. 2013. Enhancing personalized search by mining and modeling task behavior. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, page 1411–1420, New York, NY, USA. Association for Computing Machinery.

C. Xiong, Z. Dai, J. Callan, Z. Liu, and R. Power. 2017. End-to-end neural ad-hoc ranking with kernel pooling. *ACM SIGIR FORUM*, 51(cd):55–64.

Jing Yao, Zhicheng Dou, and Ji-Rong Wen. 2020a. Employing personal word embeddings for personalized search. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 1359–1368, New York, NY, USA. Association for Computing Machinery.

Jing Yao, Zhicheng Dou, Jun Xu, and Ji-Rong Wen. 2020b. Rlper: A reinforcement learning model for personalized search. In *Proceedings of The Web Conference 2020*, WWW '20, page 2298–2308, New York, NY, USA. Association for Computing Machinery.

Kun Zhou, Hui Wang, Wayne Xin Zhao, Yutao Zhu, Sirui Wang, Fuzheng Zhang, Zhongyuan Wang, and Ji-Rong Wen. 2020a. S3-rec: Self-supervised learning for sequential recommendation with mutual information maximization. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 1893–1902, New York, NY, USA. Association for Computing Machinery.

Yujia Zhou, Zhicheng Dou, and Ji-Rong Wen. 2020b. Encoding history with context-aware representation learning for personalized search. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 1111–1120, New York, NY, USA. Association for Computing Machinery.

## Appendix A Mutual Information Maximization

Mutual information maximization is a pivotal strategy for integrating diverse forms of historical information. Rooted in information theory, mutual information (MI) is a valuable tool for quantifying the dependence between random variables. Its mathematical definition is expressed as:

$$I(A,B) = P(A) - P(A|B) = P(B) - P(B|A). \tag{17}$$

Suppose $A$ and $B$ are different views of the input data, such as a word and its context in NLP tasks or a document and its historical context sequence in personalized search. Let $f$ be a function receiving $A = a$ and $B = b$ as inputs. The primary aim of maximizing MI is to tune the parameters of $f$ to maximize the mutual information $I(A,B)$, thereby extracting the most discriminative and salient attributes of the samples.

The essence of effective feature extraction involves distinguishing a sample from the entire dataset by capturing its distinctive information. By maximizing mutual information, one can isolate and harness such unique characteristics. However, when $f$ constitutes neural networks or other encoders, directly optimizing MI is usually tricky (Paninski and Liam, 2014). Thus, a common workaround is to find a tractable lower bound for $I(A,B)$ that closely approximates the target function. A specific lower bound proved to be effective in practice is InfoNCE (Logeswaran and Lee, 2018; van den Oord et al., 2018), which is based on noise contrast estimation (Gutmann and Hyvärinen, 2012). InfoNCE is defined as follows:

$$\text{InfoNCE} = \mathbb{E}_p(A,B)\left(f_\theta(a,b) - \mathbb{E}_{q(\tilde{\mathcal{B}})}\left(log\sum_{\tilde{b}\in\tilde{\mathcal{B}}}\exp f_\theta(a,\tilde{b})\right)\right) + log|\tilde{\mathcal{B}}|, \tag{18}$$

where $a$ and $b$ are different views of the input data, and $f_\theta \in \mathbb{R}$ is a function whose parameter is $\theta$ (for example, dot product result expressed by word and context or cos distance). $\tilde{\mathcal{B}}$ is a set of samples taken from the distribution $q(\tilde{\mathcal{B}})$. The $B$ set contains a positive sample $b$ and $|\tilde{\mathcal{B}}|-1$ negative samples. Learning representation based on this goal is also called contrastive learning.

We can see that the InfoNCE is analogous to the cross-entropy form the formula below when $\tilde{\mathcal{B}}$ can take all possible values of $B$ (i.e., $\tilde{\mathcal{B}} = \mathcal{B}$) and they are uniformly distributed, maximizing InfoNCE is analogous to maximize the cross-entropy loss:

$$\mathbb{E}_p(A,B) = \left(f_\theta(a,b) - \log\sum_{\tilde{b}\in\mathcal{B}}\exp f_\theta(a,\tilde{b})\right). \tag{19}$$

## Appendix B Implementation Details

The parameters of our model PSMIM are set as follows: The word embedding size is 100. The hidden size of the transformer layer in our base model is 512. The number of heads in multi-head attention is 8. The size of the MLP hidden layer is 256. In the experiment, we set the hyperparameters $z_1$, $z_2$ and $z_3$ in Formula 15 of three auxiliary tasks to 1.0. We use the Adam optimizer to minimize the final loss $L_{\text{total}}$, and the learning rate of our optimizer is $1e^{-3}$. In the experiment, we set $\alpha$ in Formula 16 to 1.0. In addition, the number of matched cores for the KRNM model is 11.

CCL 2024

# Joint Similarity Guidance Hash Coding Based on Adaptive Weight Mixing Strategy For Cross-Modal Retrieval

**Yaqi Sun** **Jing Yun** **Zhuoqun Ma**
College of Data Science and Application, Inner Mongolia University of Technology, Huhhot, China
yunjing@imut.edu.cn

## Abstract

There is a continuous and explosive growth of multimodal data. Efficient cross-modal hashing retrieval is of significant importance in conserving computational resources.To further enhance the attention to informative data within modalities and capture the semantic correlations in cross-modal data, we propose an enhanced deep Joint-Semantics Reconstructing Hashing algorithm, which is the Joint Similarity Guidance Hash Coding Based on Adaptive Weight Mixing Strategy(JSGHCA). The algorithm focuses on delving deeper into the correlations of the data in cross-modal. We introduce the adaptive weight mixing strategy to construct the semantic affinity matrix, so that the matrix can identify each modal data with specific weight in each batch. At the same time, in the process of the hash code generation, we introduce collaborative attention mechanism. It helps the model to pay more attention to the local information of each modality, thereby capturing the semantic features within each modality more accurately. Additionally, it enables the model to jointly process the attention across different modalities and extract shared semantic features more precisely. Experimental results show that the proposed model is significantly better than the deep joint semantic reconstruction hash algorithm on multiple benchmark datasets.

## 1 Introduction

With the rapid development of digital technology, multimodal information is omnipresent in our lives. On the internet, various modalities of data such as text, images, videos, and audio provide us with a rich information experience. The diverse structures of different modalities of data give rise to data heterogeneity. This heterogeneity poses a new challenge: how to effectively establish semantic associations and perform retrieval across different modalities?

Cross-modal retrieval is divided into cross-modal real-value retrieval and cross-modal hash retrieval. In cross-modal real-value retrieval, the dimensionality of real-valued vectors is relatively high, requiring large-scale matrix operations and high computational complexity. Cross-modal hash retrieval aims to eliminate differences between modalities and project them into a low-dimensional common Hamming space for retrieval tasks. In this regard, Deep Joint-Semantics Reconstructing Hashing (DJSRH (Su et al., 2019)) marks a significant milestone. It optimizes cross-modal semantic relationships by introducing a joint semantic affinity matrix. However, there is interference from irrelevant information of the two modalities during the synthesis of the joint semantic affinity matrix. Additionally, there are challenges in capturing intra-modal local information and utilizing cross-modal related information during hash code generation.

Considering these limitations, we propose a Joint Similarity Guidance Hash Coding Based on Adaptive Weight Mixing Strategy (JSGHCA) algorithm. The traditional fixed weight allocation strategy may be affected by irrelevant information when processing multimodal data, leading to a decrease in retrieval accuracy. To this end, we propose a dynamic weight allocation method that can dynamically adjust weights based on the importance of modal information within the data batch. Through this method, we

©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 999-1010, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

can more accurately identify and emphasize useful modal information when constructing a joint semantic affinity matrix, while filtering out useless information, thereby improving the accuracy and robustness of retrieval. When dealing with the local information within a single modality and capturing the complex semantic correlation between different modalities, the existing methods often have shortcomings. To address this issue, we introduce an enhanced local attention mechanism (Single Attention Mechanism, Si-Attention) that focuses more on key information within the modality during the feature extraction stage. Meanwhile, we also designed a cross modal semantic capture module (Joint Attention Mechanism, Jo-Attention) that can effectively capture and utilize the semantic associations between different modalities during the hash code generation process. Through these improvements, our model can generate richer and more accurate hash codes, thereby improving the performance of cross modal retrieval. The main contributions of this research are as follows:

- We introduce an adaptive weight mixing strategy. This approach dynamically adjusts weights based on the importance of modal information within each data batch, enhancing the retrieval process and filtering out irrelevant information.
- We introduce an attention mechanism (Collaborative Attention Mechanism) to ensure that the network primarily focuses on core information, thereby significantly enhancing the accuracy and relevance of hash encoding. We introduce an enhanced local attention mechanism Si-Attention that focuses on key information within individual modalities during the feature extraction phase. Additionally, we design a joint attention mechanism Jo-Attention to effectively capture and utilize the complex semantic correlations between different modalities in the hash code generation process.
- We conduct systematic experimental validation and performance evaluation, encompassing a series of comprehensive experiments aimed at thoroughly assessing the performance and effectiveness of JSGHCA. Through this approach, we aim to provide a more powerful, flexible, and practical hashing solution for cross-modal data, bringing new breakthroughs to the field of cross-modal information retrieval.

## 2 Related Work

Cross-modal retrieval, an extension of single-modal retrieval, tackles the intricate relationships within multimodal data. Recently, cross-modal hashing learning has been extensively employed to reduce feature storage and simplify retrieval.Cross-modal hashing learning can be divided into supervised and unsupervised methods based on the need for semantic supervision.

### 2.1 Supervised cross-modal hashing learning

Supervised cross-modal hashing learning utilizes source data and label information to explore the similarity between image and text modalities. For example, SDDH (Qin et al., 2021) employs orthogonal and equilibrium constraints to maintain modality similarity, QDCMH (Liu et al., 2021) constructs a quadruplet loss function to preserve semantic correlation, OLCH (Yi et al., 2021) acquires semantic labels through multi-class classification, NSDH (Yang et al., 2021) utilizes a three-layer network to extract features and generate hash codes, while $MSLF_1$ (Cheng et al., 2021) maximizes shared factors to obtain effective hash codes. Despite performing well on labeled cross-modal data, these methods heavily rely on large annotated datasets, thus limiting their practical applicability.

### 2.2 Unsupervised cross-modal hash learning

Unsupervised cross-modal hashing learns binary codes by leveraging modality correlations without the need for class labels. Some methods rely on constructing instance-level affinity graphs or similarity matrices to deeply analyze relationships between instances, revealing valuable similarity information. For instance, SIM (Jia et al., 2020) builds a node similarity graph, utilizing the triangle inequality principle and path averages to determine distances between nodes. SRCH (Wang et al., 2021) constructs K-nearest neighbor graphs for image and text instances, where the new graph only contains consistent relationships. UGACH (Zhang et al., 2018) creates graphs to represent relationships between image and

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 999-1010, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

text instances, typically requiring the pre-construction of the entire training set's graph structure, which is time-consuming and requires large storage space, limiting its application to large-scale data. In practical applications, such methods need to balance accuracy and computational feasibility.

Another type of unsupervised cross-modal hashing method first extracts real-valued features from different modal data and constructs a similarity matrix of these features through feature multiplication. Then, it maps the data to a hash space using a hash function and constructs a similarity matrix of hash features. In the loss function, this method compares these two similarity matrices and optimizes the hash function by minimizing the difference between them, ensuring that the generated hash codes better preserve the similarity relationships between the original data. For example, the DJSRH (Su et al., 2019) method integrates neighborhood information of multi-modal data by constructing a joint semantic correlation matrix and trains a network to generate binary codes that can reconstruct these semantic relationships to the maximum extent. DGCPN (Yu et al., 2021) extracts global similarity information by constructing a static KNN graph, combines local similarity of training batches, constructs a real-valued feature similarity matrix, then transforms it into a hash feature similarity matrix, and iteratively optimizes cross-modal hashing learning using both. DSAH (Yang et al., 2020) adopts an encoder-decoder structure to convert real-valued features of images and text into hash features, maintaining the similarity relationships between data by constructing and comparing similarity matrices of real-valued features and hash features in the loss function. Existing multi-modal retrieval methods still exhibit instability when dealing with large-scale, unbalanced, or noisy data, and many methods are limited in practical applications.

Compared to existing cross-modal retrieval methods, our approach integrates attention mechanisms, introduces a novel adaptive weighting strategy, and employs batch processing during model training, leading to performance improvements across various aspects.

## 3 Method

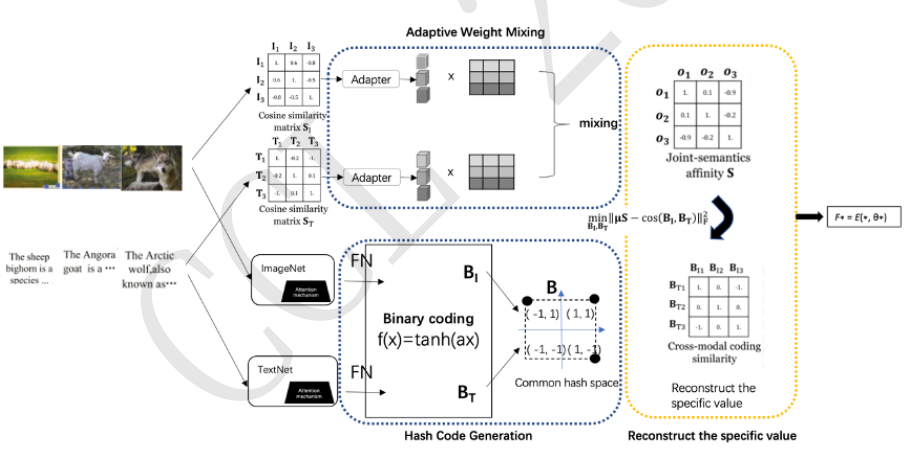



Figure 1: The structure of Cross-modal Hash Retrieval method JSGHCA

The structure of the JSGHCA method is shown in the figure1. At first, JSGHCA constructs similarity matrices for image features and text features based on the instances in each training batch. The model obtains the contribution degree of each modality in the overall similarity calculation by using the adaptive weighting strategy. We then combine the two similarity matrices according to the weights calculated within each batch, paying attention to the associations within and between modalities. In addition, JSGHCA uses pre-training network combined with collaborative attention mechanism to obtain hash features and construct hash similarity matrix, which is compared with the similar matrix of real-valued features. The model is trained by a stepwise training strategy. Firstly, the network of various modalities is trained separately, and then the cross-modal network is trained uniformly. The JSGHCA method is a comprehensive approach that combines the advantages of both deep learning and hashing techniques.

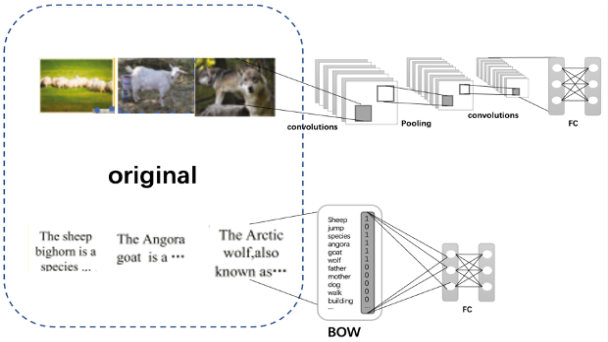



Figure 2: The extraction of the image and text features

We represent the image text pair as $p_i = (v_i, t_i)$, and the batch size is represented as m, then the whole set is $P = p_1, p_2, \cdots p_m$. For real-valued features, image and text are respectively represented as $F_V \in R^{m \times d}$ and $F_I \in R^{m \times d}$, where d represents dimension. For hash features, image and text are respectively represented as $B_V \in R^{m \times h}$ and $B_T \in R^{m \times h}$, where h represents the hash code dimension.

JSGHCA uses pre-trained model AlexNet (Krizhevsky et al., 2012) to extract image feature $F_V$, and the Bag-of-words (Gálvez-López and Tardos, 2012) model is used to extract text feature $F_T$. The formula is as follows:

$$F_* = E(*, \theta_*), * \in V, T \tag{1}$$

The hash feature length is set to a fixed value. Before converting a real-valued feature to a hash feature using function $sgn(\cdot)$, the dimension of the real-valued feature is reduced to the length of the hash code through the full join layer. The formula is as follows:

$$B_* = sgn(F_*), * \in V, T \tag{2}$$

When $sgn(\cdot)$ is used as an activation function, all non-zero inputs result in zero gradient, known as the vanishing gradient problem. To address this, we employ the scaled $\tanh(\cdot)$ function instead of $sgn(\cdot)$ because it indefinitely approximates $sgn(\cdot)$ as it approaches the limit.The formula is as follows:

$$sgn(x) = \begin{cases} 1, & \text{x} \geq 0 \\ -1, & \text{x} < 0 \end{cases} \tag{3}$$

$$\lim_{x \to \infty} \tanh(\delta x) = sgn(x) \tag{4}$$

where the parameter $\delta$ will increase with the increase of training times.

### 3.1 Adaptive weights Mixing strategy

The Unified Semantic Affinity Matrix is a two-dimensional matrix used for comparing multimodal data by mapping different data modalities into a shared semantic space. Assuming we have two modalities of feature representations $F_V$ and $F_T$, we can initially compute the similarity between them. After normalizing $F_V$, $F_T$ to $F'_V$, $F'_T$ which has unit $\ell 2 - norm$ each row, we can calculate the cosine similarity matrices $S_I = F'_I F'^T_I \in [-1, +1]^{m \times m}$ and $S_T = F'_T F'^T_T \in [-1, +1]^{m \times m}$ to describe the original neighborhood structure for the input images and texts respectively. Once we have the similarity matrix for each modal, we can calculate the adaptive weights. These weights can be calculated according to the similarity matrix of each modal, which can reflect the degree of contribution of each modal to the overall similarity calculation. Specifically, we can calculate adaptive weights using the following formula:

$$w_i = \frac{S_{irowavg}}{S_{itotalavg}} \times \frac{S_{totalavg}}{S_{totalrowavg}} \tag{5}$$

Where $w_i$ is the adaptive weight of modality I, $S_i$ is the similarity matrix of modal I, and $S_total$ is the sum of all modal similarity matrices. Through this formula, we can get the adaptive weight of each modality, which can be adjusted according to the similarity of each modality.

Algorithm1 Outlines the whole process of our proposed adaptive weight strategy.

CCL 2024

**Algorithm 1** Adaptive weight mixing strategy(AWM).

**Input:** Similarity matrix of two modalities SI and ST
**Output:** The similarity matrix after fusion FusionSimMatrix

1: Calculate the row average and total average for each modality:
$row\_avg1 = calculate\_row\_average(SimMatrix1)$
$total\_avg1 = calculate\_total\_average(row\_avg1)$
$row\_avg2 = calculate\_row\_average(SimMatrix2)$
$total\_avg2 = calculate\_total\_average(row\_avg2)$
2: Calculate the adaptive weights:
$adaptive\_weight1 = row\_avg1/total\_avg1$
$adaptive\_weight2 = row\_avg2/total\_avg2$
3: Fusion similarity matrix:
$FusionSimMatrix = zeros((len\_r(SimMatrix1), len\_r(SimMatrix2)))$
for i in range(len_r(SimMatrix1)):
for j in range(len_r(SimMatrix2)):
$FusionSimMatrix[i,j] = SimMatrix1[i,j] * adaptive\_weight1[i] + SimMatrix2[i,j] * adaptive\_weight2[i]$
4: **return** FusionSimMatrix

When we have adaptive weights for different modalities, we can utilize them to adjust and fuse the weights between modalities. By multiplying the weight of each modal by the corresponding adaptive weight, we obtain the final weight for each modal in the overall similarity calculation. Subsequently, combining the feature vectors and weights from different modalities allows us to achieve richer representations. Adjusting the weights and similarity calculation methods further optimizes the computation of the joint semantic affinity matrix, which reflects semantic similarity between modalities and guides subsequent tasks. JSGHCA introduces an adaptive weight strategy to dynamically assign weights to each modality, considering not only the inherent importance of features but also ensuring balanced representation across modalities. This approach enhances the performance of cross-modal retrieval, capturing inter-modality correlations effectively, and providing a more reliable foundation for hash coding.

### 3.2 Collaborative Attention Mechanism

The attention mechanism is used before hashing encoding forms a similarity matrix. Firstly, feature learning module is used to extract the features of image and text respectively. Considering that some information in the original data is useless for retrieval and in order to reduce the interference of these noises, this paper adds a local attention module to the network to extract the key information of each modality separately. Finally, semantic hash code is learned through semantic hash code generation module. In the process of hash code learning, joint attention mechanism is used to enhance the relationship between image and text.

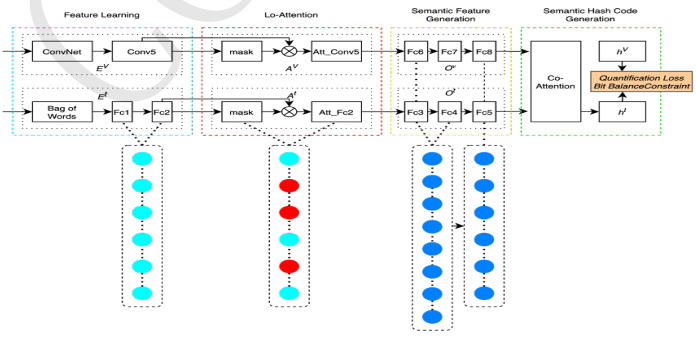



Figure 3: The structure of Collaborative Attention Mechanism

Si-Attention includes $A^V$ and $A^T$, and extracts images using an attentional mechanism of local feature extraction. For the attention operation of the image part, the image feature map $X_i^v$ obtained above is first input into the convolution layer with the convolution kernel size of $1 \times 1$ for compression, and the matrix $M_i^v = conv(X_i^v)$,$M_i^v \in R^{(H \times W)}$ is obtained. Then perform sigmoid activation function operation on $M_i^v$ to get the attention mask $\hat{M}_i^v = sigmoid(M_i^v)$,$\hat{M}_i^v \in R^{(H \times W)}$. Finally, $X_i^v$ dot the attention mask

$\hat{M}_i^v$ to get the key feature of the $i-th$ image $Y_i^v = \hat{M}_i^v \times X_i^v$,$Y_i^v \in R^{(H\times W)}$, which represents the Kronecker product. The whole process of attention processing for the $i-th$ image data can be simplified as $Y_i^v = A^v(X_i^v)$.

$$\begin{aligned} M_i^t &= fc(M_i^t) \\ \hat{M}_i^t &= sigmoid(M_i^t) \\ Y_i^t &= \hat{M}_i^t \times X_i^t \end{aligned} \tag{6}$$

For the attention operation of the text part, the key features $Y_i^t$ of the $i-th$ text feature vector $X_i^t$ can be obtained in the same way, which is implemented as follows(equation(6)):

Where fc is a fully connected layer with 8192 neurons, $M_i^t$, $\hat{M}_i^t$,$Y_i^t \in R(H \times W)$. The whole process of attention processing for the $i-th$ text feature vector can be simplified as $Y_i^t = A^t(X_i^t)$.

In order to generate more consistent hash code and reduce the information imbalance between image and text modules, a joint attention mechanism is added to this module, which uses the features learned by one module to guide the hash code training of another modal. Specifically, the generated image and text features $Q_i^v$ and $Q_i^t$are input into the joint attention mechanism, and vectors $O_i^v = fc(Q_i^v)$,$O_i^v \in R^c$ and $O_i^t = fc(Q_i^t)$,$O_i^t \in R^c$ are generated through the full connection operation with c neurons and the ReLU activation function operation. The softmax function is then used for $O_i^v$ and $O_i^t$ respectively to obtain the attention distributions $\hat{O}_i^v$ and $\hat{O}_i^t$ , i.e. $\hat{O}_i^v = softmax(Q_i^v)$, $\hat{O}_i^v \in R^c$ and $\hat{O}_i^t = softmax(Q_i^t)$, $\hat{O}_i^t \in R^c$. Finally, $Q_i^v$ and $Q_i^t$ features of the image and text are multiplied by the attention distribution $\hat{O}_i^t$ and $\hat{O}_i^v$ to update the features, respectively, and the semantic hash code $h_i^v \in R^c$ and $h_i^t \in R^c$ of the image and text is obtained, which is specifically expressed as follows:

$$\begin{aligned} h_i^v &= \hat{O}_i^t \diamond Q_i^v \\ h_i^t &= \hat{O}_i^v \diamond Q_i^t \end{aligned} \tag{7}$$

Where $\diamond$ represents the Hadamard Product. According to formula (7), it can be found that in the proposed joint attention mechanism, attention distribution $\hat{O}_i^t$ can be used to guide image semantic hash code $h_i^v$ generation, while attention distribution $\hat{O}_i^v$ can be used to guide text semantic hash code $h_i^t$ generation.

### 3.3 Joint similarity matrix guides hash coding

By constructing a joint semantic affinity matrix S, we aim to uncover the latent semantic relationships among batch input instances. This is achieved by minimizing the reconstruction error between the desired neighborhood matrix S and the hash code structure $cos(B_V, B_T)$ to be learned, thereby acquiring semantically relevant binary codes. The main functions of this approach are twofold: on the one hand, the introduction of hyperparameter $\mu$ enhances the flexibility of the reconstruction process, allowing it to better adapt to diverse data requirements; on the other hand, it constructs a specific similarity value among modalities to facilitate learning.

$$\begin{aligned} J &= \min_{B_V, B_T} \| \mu S - cos(B_V, B_T) \|_F^2 \\ s.t. S &= C(S_V, S_T) \in [-1, +1]^{(m\times m)} \end{aligned} \tag{8}$$

When we use hash coding to represent the similarity between images and texts, for instance, there are two pictures, one of a sunset beach and the other of a sunrise beach. They are actually very similar, but according to strict similarity measures, they may not be classified into the same category. The hyperparameter $\mu$ acts like a "flexibility" adjuster. When we feel that the judgment criteria are too strict, we can increase $\mu$ to relax the standards; conversely, if we want to more accurately determine similarity, we can decrease $\mu$. In this way, we can more accurately capture the true similarity relationship between images and texts.

The focus of the reconstruction framework is on maximizing the reconstruction of specific similarity values in S, rather than merely emphasizing their relative order. This approach is less sensitive to the composition of each randomly sampled training batch, making it more suitable for batch processing input methods. This article adopts the aforementioned method as the core model to establish comprehensive

CCL 2024

and systematic training objectives. In addition to integrating the multimodal reconstruction components of $B_V$ and $B_T$ in the framework, it also strengthens the internal modal reconstruction. This is because according to previous research, fully considering the internal and inter-connectivity during cross-modal network training can effectively improve retrieval efficiency. Therefore, this article utilizes the following optimized training objectives for JSGHCA.

$$J = \min_{B_V, B_T} || \mu S - cos(B_V, B_T) ||_F^2 + \lambda_1 || \mu S - cos(B_V, B_T) ||_F^2 + \lambda_2 || \mu S - cos(B_V, B_T) ||_F^2 \quad (9)$$

$\lambda_1$ and $\lambda_2$ are tradeoff parameters for balancing inter-modal reconstruction and intra-modal reconstruction. Introducing the hyperparameter $\mu$ into the framework makes the reconstruction more flexible. Compared to the Laplacian-constrained model, selecting specific similarity values for reconstruction is more compatible with batch training.

### 3.4 Training and Objective Function

Image and text belong to two different modalities, but the previous work unified the training of the data of these two modalities, ignoring their respective rules and characteristics. Therefore, the JSGHCA method adopts a step-by-step training method, firstly training the image network and text network separately, and then uniformly training the whole network, and defines the parameter $\lambda$ Used to adjust the loss function range. The overall procedure of our proposed JSGHCA is summarized in Algorithm2.

**Algorithm 2** The training process of JSGHCA method.

**Input:** image dataset I, text dataset T, batch size m, hash code length c, maximum training batch N, parameter values $\lambda$,$\mu$,...
**Output:** Feature extraction function $F* = E(*, \theta*), * \in \{I, T\}$

for $n = 1$ to N do
Extract real-valued features and hash features from image set I and text set T.
Construct the domain similarity matrix SI and ST of real-valued features.
Use adaptive weights, SI and ST to construct joint similarity matrix S.
Real-valued features are transformed into hash features through the attention mechanism.
Construct the similarity matrix BI and BT of hash features.
Construct the joint similarity matrix of hash matrix.
Calculate the objective function, backpropagate the gradient with chain rules, and update the whole parameter;
end for
**return** $F* = E(*, \theta*), * \in \{I, T\}$.

## 4 Experiments

### 4.1 Datasets

MIRFlickr (Huiskes and Lew, 2008) originates from the Flickr website, and the original MIRFLICKR 25K dataset comprises 25,000 image-text pairs, each belonging to multiple categories. In the experiment, only samples with text modalities of at least 20 words are used, resulting in a total of 20,015 image-text pairs, each belonging to 24 categories. The text is encoded into a 1386-dimensional Bag-of-Words (BOW) feature vector. The NUS-WIDE (Chua et al., 2009) dataset contains 269,648 network images, each accompanied by user-provided auxiliary tags, with each image labeled with one or more categories out of 81. Following the experimental setup in SSAH (Jin et al., 2020), image-text pairs without labels or auxiliary markers were removed, resulting in the selection of 190,421 pairs corresponding to 21 common categories. From this, 2,100 samples were randomly chosen as the query set, with the remaining 188,321 samples used as the database, and a further 10,500 samples randomly selected from the database for the training set.

| Dataset | Training set | Database | Search set | Class | Image | Text |
|---|---|---|---|---|---|---|
| MIRFLICKR-25K | 10000 | 18015 | 2000 | 24 | 4096-D | 1368-D |
| NUS-WIDE | 10500 | 188321 | 2100 | 81 | 4096-D | 1000-D |

Table 1: Statistics of the experimental dataset ("D" represents the feature dimension of the modality)

### 4.2 Evaluation Metrics

The most commonly used precision and recall rates are used as evaluation indicators in the experiment, and the corresponding top-K-precision curve can be obtained by using the obtained Precision and recall rates (abscissa: Top-k; Ordinate: precision).

The average precision AP value is the area under the precision-recall curve obtained by integrating, and the mAP value is the average AP value of all categories, calculated as follows:

$$AP(i) = \int_0^1 p(r)dr \tag{10}$$

$$mAP = \frac{\sum_{i=1}^{M} AP(i)}{M} \tag{11}$$

Where i refers to the class sequence number and M refers to the total number of classes.

### 4.3 Comparative experiment

**Experimental Details**:JSGHCA employs stochastic gradient descent optimizer with a momentum of 0.9 and weight decay of 0.0005. Evaluation during testing phase is conducted by calculating mAP values and Top-k-Precision curves. The overall evaluation criterion encompasses the mAP for image retrieval with text query and the mAP for text retrieval with image query. Parameter configuration should comprehensively consider the final experimental performance and training efficiency. After training the dataset for 10 epochs, the model demonstrates optimal training efficiency and retrieval performance. Learning rates are set to 0.001 for ImgNet and 0.01 for TxtNet, with a batch size of 32 chosen to balance performance and training efficiency. Hyperparameters are cross-validated, with $\eta = 0.4$ and $\mu$=1.5 selected. For the NUS-WIDE dataset, $\beta = 0.6$ and $\lambda_1 = \lambda_2 = 0.1$. For the MIRFlickr dataset, $\beta = 0.9$ and $\lambda 1 = \lambda 2 = 0.1$.

**Retrieval performance**:To verify the effectiveness of the JSGHCA method, it was compared with state-of-the-art cross-modal hash retrieval methods, including AAH (Fang et al., 2021), HSIDHN (Chen et al., 2021), HNH (Zhang et al., 2021), DSPH (Yun et al., 2022), FDDH (Xin et al., 2021), SDDH (Qin et al., 2021), NSDH (Yang et al., 2021), MLSPH (Zhang et al., 2021), MSDH (Zhu et al., 2021), $MSLF_2$ (Song et al., 2021), QDCMH (Liu et al., 2021) and IRGR (Li et al., 2023). The results show that JSGHCA performs well on both the MIRFlickr and NUS-WIDE datasets. Although it may not achieve the best results in single searches, combining the mAP values of both search methods, JSGHCA and HNH perform the best. Compared to other advanced methods, JSGHCA demonstrates competitive performance on large datasets, indicating its effectiveness and robustness.

| Method | I2T | | | | T2I | | | |
|---|---|---|---|---|---|---|---|---|
| | 16bits | 32bits | 64bits | 128bits | 16bits | 32bits | 64bits | 128bits |
| AAH | 0.7145 | 0.7230 | 0.7271 | 0.7283 | 0.8137 | 0.8198 | 0.8251 | 0.8281 |
| DSPH | 0.6473 | 0.6610 | 0.6703 | - | 0.6581 | 0.6781 | 0.6818 | - |
| FDDH | - | 0.7392 | 0.7578 | 0.7631 | - | 0.8022 | 0.8250 | 0.8357 |
| HNH | - | 0.8830 | 0.8950 | 0.9020 | - | 0.8540 | 0.8680 | 0.8780 |
| HSIDHN | 0.7978 | 0.8097 | 0.8179 | - | 0.7802 | 0.7946 | 0.8115 | - |
| HSIDHN | 0.7978 | 0.8097 | 0.8179 | - | 0.7802 | 0.7946 | 0.8115 | - |
| MLSPH | 0.8076 | 0.8235 | 0.8337 | - | 0.7852 | 0.8041 | 0.8146 | - |
| MSDH | 0.7836 | 0.7905 | 0.8017 | - | 0.7573 | 0.7635 | 0.7813 | 0.7959 |
| $MSLF_2$ | 0.6988 | 0.7175 | 0.7222 | 0.7294 | 0.7572 | 0.7763 | 0.7892 | 0.8229 |
| NSDH | 0.7363 | 0.7561 | 0.7656 | 0.7712 | 0.7836 | 0.8014 | 0.8183 | - |
| QDCMH | 0.7635 | 0.7688 | 0.7713 | - | 0.7762 | 0.7725 | 0.7859 | 0.8328 |
| IRGR | 0.8310 | 0.8550 | 0.8770 | 0.8940 | 0.8250 | 0.8770 | 0.8820 | 0.8920 |
| JSGHCA | 0.830 | **0.8897** | **0.9016** | 0.9012 | **0.8401** | **0.8832** | **0.9053** | **0.9102** |

Table 2: Performance comparison of various methods on MIRFlickr dataset

It should be pointed out that IRGR method ,HNH method and JSGHCA method adopt the method of constructing similarity matrix for two modalities respectively, and further process the similarity matrix on this basis, but the processing methods of the similarity matrix are different between the two methods. The experimental results also show that IRGR method, HNH method and JSGHCA method have the best

| Method | I2T | | | | T2I | | | |
|---|---|---|---|---|---|---|---|---|
| | 16bits | 32bits | 64bits | 128bits | 16bits | 32bits | 64bits | 128bits |
| AAH | 0.6409 | 0.6439 | 0.6515 | 0.6549 | 0.7379 | 0.7533 | 0.7595 | 0.7629 |
| FDDH | - | 0.6970 | 0.6910 | 0.7118 | - | 0.8133 | 0.8111 | 0.8244 |
| HNH | - | 0.8020 | 0.8160 | 0.8470 | - | 0.7760 | 0.7960 | 0.8020 |
| HSIDHN | 0.6498 | 0.6787 | 0.6834 | - | 0.6396 | 0.6529 | 0.6792 | - |
| MLSPH | 0.6405 | 0.6604 | 0.6734 | - | 0.6433 | 0.6633 | 0.6724 | - |
| MSDH | 0.6633 | 0.6859 | 0.7155 | - | 0.6359 | 0.6632 | 0.6934 | - |
| $MSLF_2$ | 0.6213 | 0.6339 | 0.6374 | 0.6482 | 0.7212 | 0.7427 | 0.7578 | 0.7765 |
| NSDH | 0.6418 | 0.6604 | 0.6732 | 0.6791 | 0.7658 | 0.7892 | 0.7939 | 0.8011 |
| SDDH | 0.6510 | 0.6564 | 0.6670 | 0.6733 | 0.7638 | 0.7790 | 0.7945 | 0.7990 |
| IRGR | 0.7560 | 0.7930 | 0.8160 | 0.8390 | 0.7500 | 0.7830 | 0.8040 | 0.8170 |
| JSGHCA | **0.7700** | **0.8203** | **0.8278** | **0.8602** | **0.7903** | **0.8357** | **0.8451** | **0.8500** |

Table 3: Performance comparison of various methods on NUSWIDE dataset

performance among all comparison methods, which indicates that the construction of similarity matrix is conducive to the improvement of performance, and also indicates that the fine-grained processing of similarity matrix by JSGHCA method is logically reasonable and significantly improves the experimental performance.

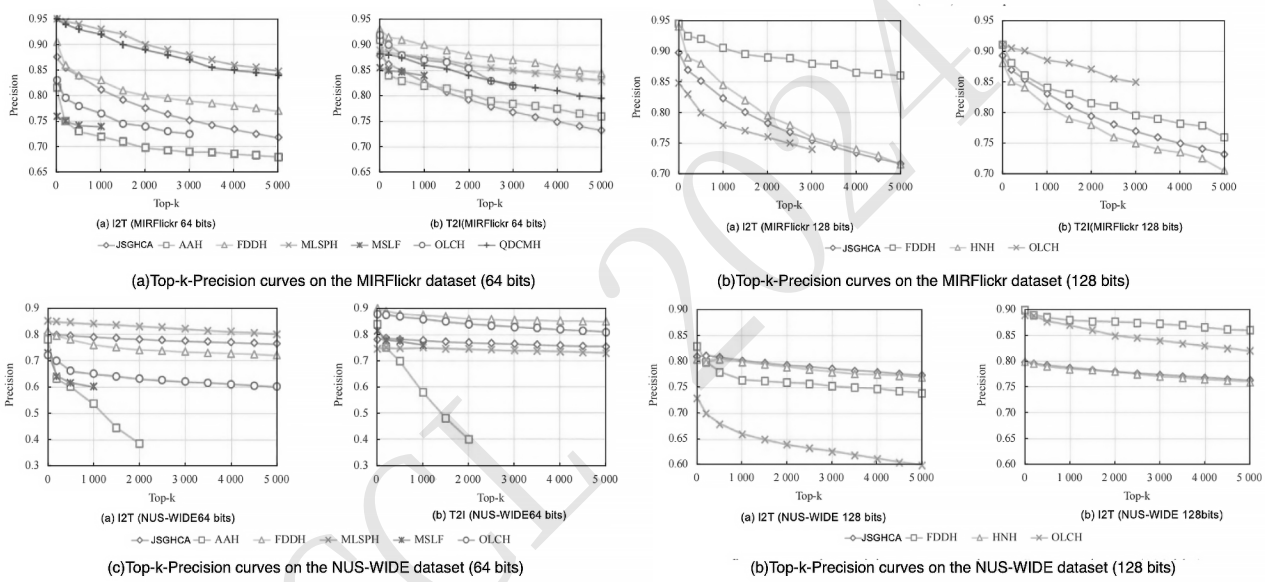



Figure 4: Top-k-Precision curves on the MIRFlickr and NUS-WIDE dataset (64 bits and 128bits)

To further compare the JSGHCA method with other cross-modal hash retrieval methods, we use a Top-k-Precision curve test. Here we compare the results of 64-bit and 128-bit hash lengths on MIRFlickr and NUS-WIDE datasets by several mainstream methods, as shown in Figure 4. As can be seen from the curve in the figure, although the accuracy of JSGHCA method on the same breakpoint may not be the highest on MIRFlickr dataset and NUS-WIDE dataset, it has relatively high performance and great advantages among similar methods, and is generally competitive.

### 4.4 Ablation Experiments

To demonstrate the effect of various innovations of the JSGHCA method, ablation experiments were performed on the data set. The ablation experiments of each module were conducted on the basis of JSGHCA, and their configurations were shown in Table 4. The JSGHCA-1 and JSGHCA-2 methods seperately used image similarity matrix and text similarity matrix to replace the joint semantic affinity matrix. The JSGHCA-3, JSGHCA-4, JSGHCA-5 and JSGHCA-6 methods do not use adaptive weight mixing strategy, Si-Attention Mechanism, Jo-Attention Mechanism and Collaborative Attention Mechanism, respectively.DJSRH is a benchmark model without adaptive weight strategy and collaborative attention Mechanism.

| modal | setting |
|---|---|
| JSGHCA-1 | $\| S_I - cos(B_I, B_T) \|_F^2$ |
| JSGHCA-2 | $\| S_T - cos(B_I, B_T) \|_F^2$ |
| JSGHCA-3 | No-AWM |
| JSGHCA-4 | No-Lo-Attention Mechanism |
| JSGHCA-5 | No-Co-Attention Mechanism |
| JSGHCA-6 | No Mechanism |

Table 4: Configuration of JSGHCA ablation experiments

Table 5-6 shows the experimental results of the ablation experiment. It can be seen from the data in the table that the JSGHCA-1 and JSGHCA-2 methods neglect the interaction and association between different modalities, focusing solely on the feature similarity within a single modality. Consequently, they may not perform well in multimodal retrieval tasks. JSGHCA-3 does not dynamically adjust weights based on the importance of different modalities, which may result in the inadequate utilization of information from various modalities.JSGHCA-4 and JSGHCA-5 methods neglect the importance of capturing crucial associations within a single modality (for JSGHCA-4) and across different modalities (for JSGHCA-5). Meanwhile, JSGHCA-6 ignores associations both within a single modality and across modalities as a whole, thus impeding its ability to fully comprehend multimodal data and generate accurate hash features.

| Method | I2T | | | | T2I | | | |
|---|---|---|---|---|---|---|---|---|
| | 16bits | 32bits | 64bits | 128bits | 16bits | 32bits | 64bits | 128bits |
| DJSRH | 0.810 | 0.843 | 0.862 | 0.876 | 0.786 | 0.822 | 0.835 | 0.847 |
| JSGHCA-1 | 0.6670 | 0.7082 | 0.7323 | 0.7745 | 0.6542 | 0.6835 | 0.7344 | 0.7632 |
| JSGHCA-2 | 0.6634 | 0.6966 | 0.7222 | 0.7422 | 0.5573 | 0.5623 | 0.6438 | 0.6532 |
| JSGHCA-3 | 0.8122 | 0.8623 | 0.8755 | 0.8802 | 0.8113 | 0.8582 | 0.8644 | 0.8723 |
| JSGHCA-4 | 0.8186 | 0.8732 | 0.8911 | 0.8887 | 0.8096 | 0.8654 | 0.8873 | 0.8912 |
| JSGHCA-5 | 0.8108 | 0.8555 | 0.8700 | 0.8864 | 0.8092 | 0.8487 | 0.8673 | 0.8774 |
| JSGHCA-6 | 0.7902 | 0.8538 | 0.8792 | 0.8805 | 0.7756 | 0.8324 | 0.8606 | 0.8762 |
| JSGHCA | **0.8300** | **0.8897** | **0.9016** | **0.9012** | **0.8401** | **0.8832** | **0.9053** | **0.9102** |

Table 5: Comparison results on the MIRFlickr dataset

| Method | I2T | | | | T2I | | | |
|---|---|---|---|---|---|---|---|---|
| | 16bits | 32bits | 64bits | 128bits | 16bits | 32bits | 64bits | 128bits |
| DJSRH | 0.724 | 0.773 | 0.798 | 0.817 | 0.712 | 0.744 | 0.771 | 0.789 |
| DJSRH$_{S=S_I}$ | - | - | 0.717 | 0.741 | - | - | 0.712 | 0.735 |
| DJSRH$_{S=S_T}$ | - | - | 0.702 | 0.734 | - | - | 0.606 | 0.581 |
| JSGHCA-1 | 0.6753 | 0.6952 | 0.7265 | 0.7803 | 0.7098 | 0.7320 | 0.7696 | 0.7754 |
| JSGHCA-2 | 0.6792 | 0.6866 | 0.7024 | 0.7865 | 0.7521 | 0.7542 | 0.7853 | 0.8042 |
| JSGHCA-3 | 0.7423 | 0.7855 | 0.8025 | 0.8244 | 0.7652 | 0.7843 | 0.8413 | 0.8424 |
| JSGHCA-4 | 0.7322 | 0.7753 | 0.7902 | 0.8265 | 0.7756 | 0.8324 | 0.8366 | 0.8382 |
| JSGHCA-5 | 0.7532 | 0.7462 | 0.8321 | 0.8902 | 0.7653 | 0.8242 | 0.8234 | 0.8253 |
| JSGHCA-6 | 0.7676 | 0.8032 | 0.8153 | 0.8439 | 0.7522 | 0.7900 | 0.8102 | 0.8199 |
| JSGHCA | **0.7700** | **0.8203** | **0.8278** | **0.8602** | **0.7903** | **0.8357** | **0.8451** | **0.8500** |

Table 6: Comparison results on the NUSWIDE dataset

To sum up , the shortcomings of these ablation experiments suggest that various innovations in JSGHCA are critical to improving the performance and accuracy of multimodal retrieval tasks. By considering the interaction and association of different modalities, using adaptive weight mixing strategies, and introducing Si-Attention, Jo-Attention, and collaborative-attention mechanisms, JSGHCA is able to better understand multimodal data and generate accurate hash features, thus achieving excellent performance in multimodal retrieval tasks.

## 5 Conclusion

In this paper, we introduced an improved deep joint semantics reconstructing hashing algorithm called JSGHCA. This algorithm combines adaptive weight strategies and attention mechanisms to enhance the

accuracy and descriptiveness of hash encodings for image and text modalities. Experimental results demonstrate that JSGHCA outperforms existing algorithms in multi-modal retrieval tasks, particularly with the incorporation of attention mechanisms and adaptive weight strategies. Additionally, we believe there is further room for exploration, such as integrating more advanced pre-trained models or optimizing attention mechanisms. We look forward to seeing the application of JSGHCA in other tasks, such as image and text generation, multi-modal translation, and beyond. JSGHCA opens up new possibilities in multi-modal retrieval and sets a new benchmark for future research.

## Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62062055, Inner Mongolia University Youth Science and Technology Talent under Grant NJYT24061 and the basic Scientific Research Services of Colleges and Universities directly under Inner Mongolia Autonomous under Grant JY20220249.

## References

Chen S, Wu S., Wang L. Hierarchical semantic interaction-based deep hashing network for cross-modal retrieval. PeerJ. Computer science, 2021, 7:e552.

Cheng X, Liu Z, Zhang Q. MSLF: Multi-scale legibility function to estimate the legible scale of individual line features. Cartography and Geographic Information Science, 2021, 48(2), 151-168.

Chua T S, Tang J, Hong R, et al. NUS-wide: a real-world web image database from National University of Singapore. Proceedings of the ACM international conference on image and video retrieval, 2009, pp. 1-9.

Fang X, Jiang K, Han N, et al. Average Approximate Hashing-Based Double Projections Learning for Cross-Modal Retrieval. IEEE Transactions on Cybernetics, 2021,PP(99), pp. 1-14.

Gálvez-López D, Tardos J D. Bags of binary words for fast place recognition in image sequences. IEEE Transactions on robotics, 2012, 28(5), pp. 1188-1197.

Huiskes M J, Lew M S. The mirflickr retrieval evaluation. Proceedings of the 1st ACM international conference on Multimedia information retrieval, 2008, pp. 39-43.

Jia M, Zhai Y, Lu S, et al. A similarity inference metric for RGB-infrared cross-modality person re-identification. arXiv preprint arXiv:2007.01504, 2020.

Jin S, Zhou S, Liu Y, et al. SSAH: Semi-supervised adversarial deep hashing with self-paced hard sample generation. Proceedings of the AAAI conference on artificial intelligence, 2020, 34(07), pp. 11157-11164.

Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 2012, 25.

Li Z X, Hou C W, Xie X M. Unsupervised Cross-modal Hash Retrieval Fusing Multiple Instance Relations. Ruan JianXue Bao/Journal of Software, 2023, 34(11), pp. 4973–4988 (in Chinese).

Liu H, Xiong J, Zhang N, et al. Quadruplet-based deep cross-modal hashing. Computational Intelligence and Neuroscience, 2021, 2021.

Qin J, Fei L, Zhu J, et al. Scalable Discriminative Discrete Hashing For Large-Scale Cross-Modal Retrieval. ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021.

Song W, Huan Z, Kei N. Learning a maximized shared latent factor for cross-modal hashing. Knowledge-Based Systems, 2021, 228.

Su Shupeng, Zhisheng Zhong, and Chao Zhang. Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval. Proceedings of the IEEE/CVF International Conference on Computer Vision. 3027-3035, 2019.

Wang W, Shen Y, Zhang H, et al. Set and rebase: Determining the semantic graph connectivity for unsupervised cross-modal hashing. Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence, 2021, pp. 853-859.

CCL 2024

Xin L, Xingzhi W, YiuMing C. FDDH: Fast Discriminative Discrete Hashing for Large-Scale Cross-Modal Retrieval. IEEE transactions on neural networks and learning systems, 2021.

Yang D, Wu D, Zhang W, et al. Deep semantic-alignment hashing for unsupervised cross-modal retrieval. Proceedings of the 2020 international conference on multimedia retrieval, 2020, pp. 44-52.

Yang Z, Yang L, Raymond, O I, et al. NSDH: A nonlinear supervised discrete hashing framework for large-scale cross-modal retrieval. Knowledge-Based Systems, 2021, 217, 106818.

Yi J, Liu X, Cheung Y, et al. Efficient online label consistent hashing for large-scale cross-modal retrieval. 2021 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2021, pp. 1-6.

Yu J, Zhou H, Zhan Y, et al. Deep graph-neighbor coherence preserving network for unsupervised cross-modal hashing. Proceedings of the AAAI conference on artificial intelligence, 2021, 35(5), pp. 4626-4634.

Yun L, Shujuan J, Qiang F, et al. An efficient dual semantic preserving hashing for cross-modal retrieval. Neurocomputing, 2022, 492, pp. 264-277.

Zhang J, Peng Y, Yuan M. Unsupervised generative adversarial cross-modal hashing. Proceedings of the AAAI conference on artificial intelligence, 2018, 32(1).

Zhang P, Luo Y, Huang Z, et al. High-order nonlocal Hashing for unsupervised cross-modal retrieval. World Wide Web, 2021, 24(2), pp. 1-21.

Zhang X, Wang X, E, M, et al. Multi-label semantics preserving based deep cross-modal hashing. Signal Processing: Image Communication, 2021, 93.

Zhu L, Tian G, Wang B, et al. Multi-attention based semantic deep hashing for cross-modal retrieval. Applied Intelligence, 2021, 51(8), pp. 1-13.

CCL 2024

# Generate-then-Revise: An Effective Synthetic Training Data Generation framework for Event Detection

**Huidong Du, Hao Sun, Pengyuan Liu, Dong Yu**
School of Information Science, Beijing Language and Culture University
Beijing, China
yzqtdu@gmail.com, 202221198129@stu.blcu.edu.cn
liupengyuan@pku.edu.cn, yudong_blcu@126.com

## Abstract

Large language models (LLMs) struggle with event detection (ED) due to the structured and variable number of events in the output. Existing supervised approaches rely on a large amount of manually annotated corpora, facing challenges in practice when event types are diverse and the annotated data is scarce. We propose Generate-then-Revise (GtR), a framework that leverages LLMs in the opposite direction to address these challenges in ED. GtR utilizes an LLM to generate high-quality training data in three stages, including a novel data revision step to minimize noise in the synthetic data. The generated data is then used to train a smaller model for evaluation. Our approach demonstrates significant improvements on the low-resource ED. We further analyze the generated data, highlighting the potential of synthetic data generation for enhancing ED performance.

## 1 Introduction

Event detection (ED) aims to extract structured information from natural language text. Previous researches typically involve supervised fine-tuning on a large amount of manually annotated corpora, facing challenges in practice due to the vast number of event types in real-world scenarios and the high cost of manually annotating data. Recently, large language models (LLMs) like ChatGPT and Llama2 have made remarkable progress in many tasks in low-resource scenarios, enabling some tasks to achieve performance comparable to fully supervised learning in zero-shot or few-shot conditions. However, the success of large models in natural language understanding and reasoning tasks is difficult to directly replicate in ED (Ma et al., 2023b; Qin et al., 2023; Han et al., 2023; Li et al., 2023).

Current research indicates that directly using prompt instructions to instruct models to perform ED generally results in poor performance (Huang et al., 2024). As shown in Figure 1, researchers evaluate the average performance of several LLMs on ED in a test set containing 250 documents. Among them, Mixtral-8x7B-Instruct-v0.1 achieves the best trigger identification (TI) F1 score of 37.5 in a 64-shot setting, while for trigger classification (TC), the performance of various models is even worse, and there is no linear increase in performance with the increase in samples.The failure of LLMs in ED is related to the characteristics of the task itself. ED requires generating a structured list of events of varying lengths. Although these models are not adept at structured output tasks, recent research has found that models can generate meaningful inputs based on outputs in the opposite direction (Josifoski et al., 2023), effectively transforming the original computation from $\mathcal{P}(y_k|x)$ to $\mathcal{P}(x|y_k)$.

This suggests that we can leverage the asymmetry to synthesize high-quality data for ED. Data generation avoids the difficulties of outputting structured content while reducing reliance on manually annotated data. However applying LLMs to data generation on ED is non-trivial. Unlike classification tasks like sentiment analysis (which use sentence-level labels like positive or negative) (Ye et al., 2022a; Ye et al., 2022b), ED requires generating data with token-level labels (Gao et al., 2022), i.e, each sentence should contain both the correct triggers and event types. Simply prompting LLMs can lead to noisy data, as

©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License
This work is funded by the Humanity and Social Science Youth foundation of Ministry of Education (23YJAZH184) and the Fundamental Research Funds for the Central Universities in BLCU (No. 21PT04)

| Model | TI | TC |
|---|---|---|
| Mixtral-8x7B-Instruct-v0.1 (2-shot) | 30.4 | 10.2 |
| Mixtral-8x7B-Instruct-v0.1 (6-shot) | 34.4 | 10.6 |
| Mixtral-8x7B-Instruct-v0.1 (16-shot) | 35.4 | 12.1 |
| Mixtral-8x7B-Instruct-v0.1 (32-shot) | 36.7 | 13.8 |
| Mixtral-8x7B-Instruct-v0.1 (64-shot) | 37.5 | 14.6 |
| gpt-3.5-turbo-1106 (2-shot) | 33.9 | 11.8 |
| gpt-3.5-turbo-1106 (16-shot) | 35.2 | 12.3 |
| Llama-2-70b-chat-hf (2-shot) | 30.6 | 11.3 |
| Llama-2-70b-chat-hf (6-shot) | 32.2 | 12.4 |

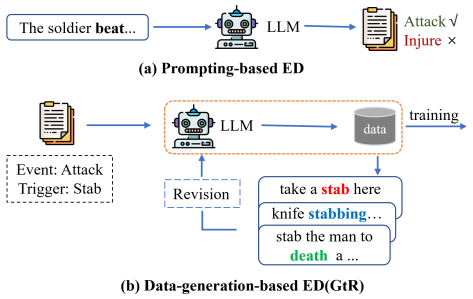



Figure 1: Table of the performance of LLms on ED and Illustrations of ED with LLMs. Compared to prompting, GtR generate synthetic data with LLMs for training. However, data generation may encounter several issues in ED. In (b), the red word represents an incorrect event, blue word suggets competing events (a same trigger for *Attack* and *Injure* event), and the green word represents an accompanying event *Death*. GtR proposes a revision step to address these problems.

shown in Figure 1, including accompanying events, competing events and incorrect events. Unlabled accompanying events emerge frequently in sythesized data and can disrupt the training process.

Existing works leveraging LLMs to generate labeled data in ED often overlook these challenges, primarily relying on sophisticated prompts to produce training data. Ma's approach (Ma et al., 2023a) leverages LLMs for data augmentation but fails to address the noise issues discussed earlier. While existing research offers noise mitigation techniques for classification tasks (e.g., noise-robust loss functions) (Ye et al., 2022b; Gao et al., 2023), these methods are ineffective against the specific problem of accompanying events in ED. To address this gap, we identify three key challenges to overcome when mitigating noise in labeled data generation: (1) Aligned Generation: The LLM must generate data that strictly adheres to the specified triggers and event types. (2) Internal Consistency: The triggers and event types within each generated sample need to be consistent and mutually supportive. (3) Accompanying Event Handling: We need a mechanism to identify and label accompanying events during generation.

To address these issues, we propose a sentence-level event data generation framework called GtR. GtR synthesizes high-quality automatically annotated data in three stages, which can be used in zero-shot scenarios without manual annotation and can also enhance existing data in few-shot settings. The synthesized data also includes adversarial examples, which can significantly improve the robustness of the model. In addition, considering that synthesized data inevitably contains some noise, we introduce a dynamic weighting module to score the synthesized data, reducing the interference of noisy data and improving the performance of models trained on synthesized data.

## 2 Related works

**Zero-Shot Event Detection** Zero-shot event detection (ED) aims to identify triggers and event types without accessing training data (Du and Cardie, 2020; Lu et al., 2022; Hsu et al., 2022). Previous approaches for zero-shot ED rely on external tools such as AMR and FrameNet to learn event features from seen events, and apply them to unseen events (Huang et al., 2018; Zhang et al., 2022b). These methods assume that seen and unseen events have similarities, which may not always hold true in real-world scenarios.

Recently, applying LLMs to zero-shot tasks has gained popularity . Compared to fully supervised learning, LLMs alleviate the trouble of limited data. However, when it comes to ED, the prompt-based method still yields relatively unsatisfactory performance (Ma et al., 2023b; Qin et al., 2023; Han et al., 2023; Li et al., 2023).

**Dataset-generation-based Zero-shot Learning** Some researches have explored an alternative approach leveraging the knowledge of LLMs for zero-shot learning. These methods prompt the LLMs to generate samples for downstream tasks based on the given task descriptions and labels (Schick and Schütze, 2021). With the generated samples, a smaller model with faster speed and lower computational

requirements can be trained. These methods have shown significant results in classification tasks. Ye et al. (Ye et al., 2022a) indicates that the synthetic data generated by LLMs is of high-quality in sentiment classification and natural language inference, even surpassing the performance achieved by fine-tuning on original data on some datasets.

Some studies recently explore data generation with LLMs in the realm of information extraction. Josifoski (2023) generates high-quality structured data using LLMs in relation extraction. Ma (2023a) applies data generation in event extraction but fails to address the issues of accompanying and incorrect events. In contrast to these methods, GtR generates samples from scratch, and proposes a revision stage to address the noise, obtaining data that can be directly used for training.

## 3 Data Generation Framework

In this section, we introduce how GtR generates high-quality labeled training data for ED, i.e. $\mathcal{S}_{syn} = \{\mathcal{X}_{sen}, \mathcal{Y}_{event}\}$.

### 3.1 Event Template

We define an event template formatted as a JSON array to specify key information about events to guide the following stages. The event template $\mathcal{E} \sim \{e_1, e_2..., e_k\}$, where $k$ is the number of events, consists of two items for every elements.

- **Description**. *Description* is the concrete definition of an event type. For example, the *Injure* event can be defined as *a person experiences physical harm.* *Description* should be precise to avoid ambiguous results when generating triggers and sentences.
- **Examples**. *Examples* provide demonstrations and CoT rationales for generating triggers and sentences, facilitating in-context learning and enhancing their overall quality. Note that we **do not** use examples in zero-shot setting for fair comparison.

### 3.2 Trigger Expansion

With the guide of an event template, we iteratively leverage LLMs to produce triggers for every event. Formally, given the model $\mathcal{M}$ and a function $\mathcal{F}_{tri} : \mathcal{E} \rightarrow \mathcal{P}_{tri}$, where $\mathcal{P}_{tri}$ denotes the prompt, we obtain generated triggers $\mathcal{T}_{tri} \sim \mathcal{M}(\mathcal{F}_{tri}(\mathcal{E}))$. $\mathcal{F}_{tri}$ retrieves event description and examples from the template, and combines them with an instruction to form the prompt fed to LLMs. The input prompt for *Injure* would be: *list 10 words which mean that a person experiences physical harm.*

### 3.3 Data Generation

Both triggers and sentences can be generated simultaneously, but in order to get better control over the type and diversity of sentences, we split them into two stages. With the function $\mathcal{F}_{sen} : (\mathcal{E}, \mathcal{T}_{tri}) \rightarrow \mathcal{P}_{sen}$ , we gain positive and negative samples $(\mathcal{X}_{sen}, \mathcal{Y}'_{event}) \sim \mathcal{M}(\mathcal{F}_{sen}(\mathcal{E}, \mathcal{T}_{tri}))$, where $\mathcal{P}_{sen}$ denotes the prompt for generating sentences, $\mathcal{X}_{sen}$ denotes the generated sentences and $\mathcal{Y}'_{event}$ represents the events. For every trigger in $T_{tri}$, $\mathcal{F}_{sen}$ combines description and examples from the template with an instruction to prompt LLMs to output sentences.

The complete prompt for positive and negative samples are nearly identical. For example, given the trigger **kill**, we build the following prompt: *make 5 sentences with the word kill. The sentences should have different backgrounds. The length and structure should vary from each other. The word should* ***mean/not mean*** *causing the death of a person*. LLMs are capable of effectively generating sentences that meet the requirements of the prompts.

### 3.4 Data Revision

Data obtained from the *Data Generation* step are highly likely noisy, containing incorrect events and accompanying events, e.g., a sentence for *Attack* may also include the *Injure* or *Die* event. Hence, we propose the *Data Revision* step to solve these problems. With $\mathcal{F}_{rev} : (\mathcal{E}, \mathcal{T}_{tri}, \mathcal{X}_{sen}, \mathcal{Y}'_{event}) \rightarrow (\mathcal{X}_{sen}, \mathcal{Y}_{event})$, we get the final training dataset.

CCL 2024

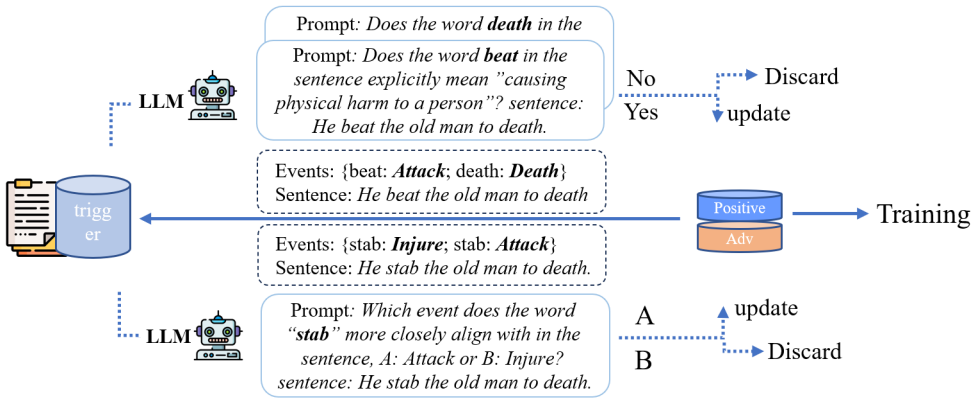



Figure 2: Data Revision: we address incorrect events, accompanying events and competing events based on the inference of LLM. As depicted in the figure, the LLM determines that event A is more appropriate than event B within current sentence.

**Incorrect and Accompanying Events** $\mathcal{F}_{rev}$ identifies events within generated text from $\mathcal{X}_{sen}$ using triggers in $\mathcal{T}_{tri}$. It first labels all words or phrases from $\mathcal{T}_{tri}$ as potential events within the sentence. Then $\mathcal{F}_{rev}$ constructs prompts leveraging a LLM to answer Yes/No questions about each potential event. For example, consider the trigger **beat** associated with the *Attack* event and the sentence *He beat the old man to death*. As shown in Figure 2, $\mathcal{F}_{rev}$ labels the word **death** as a potential *Death* event and then builds prompts for *Attack* and *Death* successively. The complete prompt is like *sentence: He beat the old man to death. Does the word beat in the sentence explicitly mean "causing physical harm to a person"?* Events receiving a **Yes** answer are retained, while those receiving a **No** are discarded. This process helps resolve incorrect and accompanying events within the generated text.

**Competing Events** We need to settle competing events in the samples. In some cases, LLMs mistakenly assign the same trigger in a sentence to different similar events, leading to race conditions. Given the trigger **stab**, LLMs may consider it both as an *Injure* and *Attack* event. To tackle this, we request LLM to explain the difference between these two events. Provided with the explanation as context, LLM determines which event is more accurate within the current sentence. Finally, we modify samples based on the answers.

After *Data Revision*, we gain the final synthetic data $\mathcal{S}_{syn}$ for training.

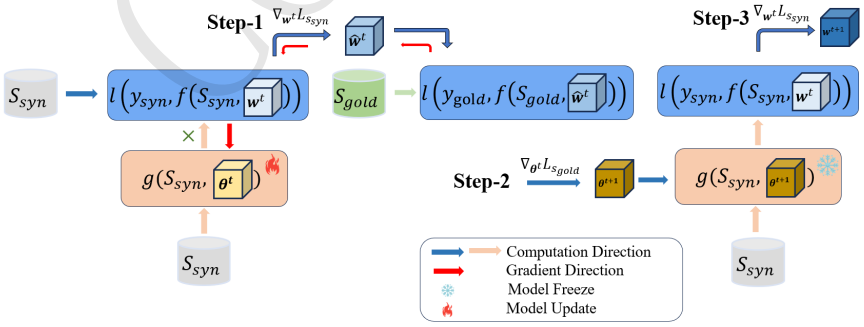



Figure 3: Illustration of the process of weighting.

## 4 Dynamic Weighting

*Data Revision* can mitigate the aforementioned noise issues. As the scale of generated data increases, residual noise will inevitably continue to grow, constraining further improvements in model performance. Therefore, we also propose a dynamic weighting module $g(x, \boldsymbol{\theta})$ that assigns weight coefficients to

samples in $\mathcal{S}_{syn}$.

### 4.1 Weighting Model

We first utilize pre-trained language models such as BERT to encode the synthesized data. Given a sample sentence $W = \{w_1, w_2, ..., w_n\}$ of length $n$, we obtain its encoded feature representation $\hat{W} = \{\hat{w_1}, \hat{w_2}, ..., \hat{w_n}\}$. The sentence contains a list of events, with each event recording a specific trigger $w_i$ and event type $e_{w_i}$. Next, we initialize learnable event type vectors $E = \{e_1, e_2, ..., e_k\}$, where $k$ is the number of event types. We score the similarity between $w_i$ and $e_{w_i}$ using $g(x, \boldsymbol{\theta})$ as follows:

$$g(w_i) = Norm(Similarity(MLP(\hat{w_i}), e_{w_i})) \tag{1}$$

Here, the similarity is computed using cosine similarity, and the similarity scores are normalized. $g(x, \boldsymbol{\theta})$ outputs scores ranging from 0 to 1 for each trigger in the event list, determining the likelihood of triggering the corresponding event

### 4.2 Model Optimization

In this subsection, we will detail the optimization process of $g(x, \boldsymbol{\theta})$. We follow Shu (2019) to optimize the model based on $\mathcal{S}_{syn}$ (Support Set) and a small set of gold data $\mathcal{S}_{gold}$ (Query Set) obtained from examples. A classification network $f(x, \boldsymbol{\omega})$ is first trained based on the Support set, where $M \in \mathcal{R}^{H \times k+1}$, $k+1$ equals the total number of event types and Null event.

$$f(x) = SoftMax(Encoder(x)M) \tag{2}$$

Next, we use the following loss function to obtain the updated parameters of the classification network $\hat{\boldsymbol{\omega}^{\boldsymbol{t}}}$ at the time of $t$:

$$L_{s_{syn}} = \frac{1}{m} \sum_{i=1}^{m} Score(w_i) L^i_{s_{syn}} \tag{3}$$

$$Score(w_i) = \begin{cases} 1, & if\ \exists\ e_{w_i} \\ g(w_i), & if\ \nexists\ e_{w_i} \end{cases} \tag{4}$$

$$L^i_{s_{syn}} = l(y^i_{s_{syn}}, f(s^i_{syn}, \boldsymbol{\omega})) \tag{5}$$

$s$ denotes the length of all tokens in $S_{syn}$ and $l$ denotes cross-entropy loss function. The loss coefficient is determined based on the return value of the score function. If a token does not trigger an event, the coefficient is defaulted to 1. If it is a trigger, the coefficient is calculated based on the weighting model. We gain $\hat{\boldsymbol{\omega}^{\boldsymbol{t}}}$ as follows:

$$\hat{\boldsymbol{\omega}^{\boldsymbol{t}}} = \boldsymbol{\omega^t} - \alpha \times \nabla_{\boldsymbol{\omega^t}} L_{s_{syn}} \tag{6}$$

At time $t+1$, we optimize the weighting model on the query set using $\hat{\boldsymbol{\omega}^{\boldsymbol{t}}}$ (Step-1 in Figure 3). Specifically, we compute the loss on the Query set, with the equations as follows:

$$L_{s_{gold}} = \frac{1}{k} \sum_{i=1}^{k} L^i_{s_{gold}} \tag{7}$$

$$L^i_{s_{gold}} = l(y^i_{s_{gold}}, f(s^i_{gold}, \hat{\boldsymbol{\omega}^{\boldsymbol{t}}})) \tag{8}$$

$k$ is the length of the Query set tokens. Based on this, we calculate the gradients of the parameters of $g(x, \boldsymbol{\theta})$ and update the parameters from $\boldsymbol{\theta^t}$ to $\boldsymbol{\theta^{t+1}}$ (Step-2 in Figure 3). Then we freeze $\boldsymbol{\theta^{t+1}}$ and calculate the gradients of the classification network's parameters at time $t+1$ according to the loss function in 3. And finally update $\boldsymbol{\omega^t}$ to $\boldsymbol{\omega^{t+1}}$ (Step-3 in Figure 3)

CCL 2024

## 5 Event Detection Model Training

After completing the learning process of the weight model, we use $g(x, \boldsymbol{\theta})$ to calculate the weights for generated data. During training, we merge the weighted synthetic data with the example data, adjust the loss coefficients based on the weights.

## 6 Experiments

**Datasets** We perform experiments on three different ED datasets including ACE05 (Walker et al., 2006), MAVEN (Wang et al., 2020) and Commodities News (Lee et al., 2021), which contain 33, 168 and 18 event types respectively. We refer to previous works to partition the test set from ACE05 (Du and Cardie, 2020) and Commodities News (Lee et al., 2021) for evaluating. For MAVEN, we use the officially provided validation set for testing.

**Baselines** We train a model without weighting on synthetic data and compare it with these baselines in zero-shot setting: UIE(Lu et al., 2022), ZEOP(Zhang et al., 2022b) ZED(Zhang et al., 2022a) , gpt-3.5-turbo and GtR_prompt. Compared with GtR, GtR_prompt omits the *Data Generation* stage and treats the targeted test set as unlabeled generated data (Generated triggers are used in this stage). Then GtR_prompt regards the labeled data obtained through the *Data Revision* stage as final results.

We train a model with weighting and compare the model with these baselines in few-shot setting: UIE (Lu et al., 2022), STAR(Ma et al., 2023a), gpt-3.5-turbo.

**Evaluation Strategy** We adopt the same evaluation metrics as previous works on ED. For trigger identification (TI), a predicted trigger is considered correct only if it matches the gold trigger exactly. For classification (TC), both triggers and types should be the same with the gold labels. We use F1 score as the evaluation metrice.

**Implementation Details** We use OpenAI's gpt-3.5-turbo for datasets generation. When constructing event templates, we refer to the official guideline for ACE05, and write event descriptions for other datasets on our own. For fair comparison, we do not use any examples but rely solely on the basic prompt in zero-shot setting. We generate 821 triggers and 6,387 sentences for ACE05, 1,647 triggers and 16,876 sentences for MAVEN, 447 triggers and 6,877 sentences for Commodities News.

We employ the pre-trained model Bert-base-uncased from the Hugging-face community[0] and use the model architecture from Du (2020) for training. We use the Adam optimizer with a learning rate of $4e^{-5}$ and set other hyper-parameters as Du (2020). All models are optimized on a single NVIDIA V100 GPU.

| Model | ACE05 | | MAVEN | | Commodities News | |
|---|---|---|---|---|---|---|
| | TI | TC | TI | TC | TI | TC |
| UIE | - | 38.14 | - | 24.17 | - | 35.64 |
| ZEOP | - | 45.91 | - | - | - | - |
| ZED | - | - | 49.39 | 32.96 | - | - |
| gpt-3.5-turbo | 24.63 | 18.75 | 25.76 | 20.32 | 23.40 | 19.76 |
| GtR_prompt | 43.92 | 41.93 | 30.12 | 28.54 | 56.23 | 34.12 |
| STAR | 43.74 | 38.90 | - | - | - | - |
| GtR | 55.29 | 50.49 | 44.60 | 36.48 | 68.79 | 45.03 |

Table 1: The zero-shot performance evaluated on three datasets. We adopt the results of UIE on ACE05 from the original paper, and conduct experiments on the other two datasets. For prompting, we employ the same model, gpt-turbo-3.5, perform three rounds of experiments and use the same set of generated triggers. Regarding GtR, we report the average results of three experiments conducted with different random seeds on the same generated data.

[0]https://huggingface.co/bert-base-uncased

### 6.1 Main Results on Zero-Shot ED

We experimentally evaluate the generative performance of GtR in a zero-shot scenario, where no examples are used, allowing for a direct validation of the framework's generative capabilities. From the table 1, we can draw the following conclusion:

1. The model trained on synthetic data consistently outperforms LLMs and demonstrates a more stable performance. We argue that GtR, through its three-stage generation process and iterative refinement, transforms the challenging task of extracting complex structured events into a series of simpler and more answerable questions, which allows for better and more consistent performance.

2. GtR_prompt performs better than directly prompting, but still falls short of the model trained on GtR. This performance gap may arise from two reasons. Firstly, GtR includes adversarial samples, which can provide additional information and reduce the risk of the model being overly confident. We validate the effectiveness of adversarial samples in subsequent experiments. Secondly, the sentences encoded by the pretrained models contain certain features, such as part-of-speech features, which can enhance ED performance.

3. GtR outperforms UIE, indicating that the advantages gained by UIE through multitask pretraining do not surpass those achieved by synthesizing data using GtR. Additionally, other methods for zero-shot ed using external knowledge and contrastive learning have not shown significant superiority over GtR.

| Model | TI | | TC | |
|---|---|---|---|---|
| | 5-shot | 10-shot | 5-shot | 10-shot |
| UIE | - | - | 51.21 | 53.23 |
| gpt-3.5-turbo | 34.22 | 34.64 | 11.17 | 12.54 |
| STAR | 61.39 | 63.57 | 56.41 | 59.10 |
| GtR | 63.47 | 65.78 | 58.82 | 61.21 |
| w/o weighting | 61.56 | 64.21 | 56.74 | 58.43 |

Table 2: Experiment results on few-shot ED

### 6.2 Main Results on Few-Shot ED

We also conduct experiments in the low-resource scenario of the ACE05 dataset, specifically in 5-shot and 10-shot settings.

1. From table 2, we find that even with an increase in examples, directly prompting LLMs did not result in linear growth in benefits. Performance on both TI and TC are comparatively poor, far behind supervised methods. Finding better prompting methods to improve the performance of LLMs in ED remains a topic worthy of attention.

2. GtR outperforms STAR, validating the superiority of our framework and the necessity of revising data. Compared to STAR, our proposed framework has a higher ceiling. It is noteworthy that both GtR and STAR outperform UIE, once again confirming that data synthesis can quickly improve the performance of ED.

3. Dynamic weighting improves the quality of synthetic data and leads to better performance. This indicates that synthetic data itself contains some noise, which becomes more constraining to the model as the scale increases. It also underscores the necessity of dynamically weighting synthetic data.

### 6.3 Ablations

Compared to other generation methods, GtR primarily introduces the *Data Revision* step for synthetic data. To explore the improvement in data quality through *Data Revision*, we conduct ablation studies on the generated datasets. As shown in Table 3, we first perform a base experiment on the basic dataset obtained during *Data Generation*. The column *Events* shows the number of events in each dataset. +Rev_old denotes the updated dataset in which only existing labels are revised, +Rev_new denotes the dataset based on +Rev_old in which accompanying events are further identified , and +Rev_com is the dataset based on +Rev_new where competing events are resolved.

CCL 2024

| | Events | P | R | F1 |
|---|---|---|---|---|
| Basic data | 3,980 | 32.47 | 24.69 | 28.05 |
| +Rev_old | -626 | 36.95 | 22.72 | 28.13 |
| +Rev_new | +1,828 | 52.08 | 46.42 | **49.09** |
| +Rev_com | -77 | 58.88 | 44.20 | **50.49** |

Table 3: Experiments conducted on different data generated during the revision phase, where the column **Events** indicates the number of events in the current dataset. Every dataset is generated based on the previous dataset.

From Table 3, we observe that model trained on the basic data performs averagely. Basic dataset is filled with noise, which hampers the model's performance. In +Rev_old, the number of events decreases, while the F1 score remains comparable to before with an improvement in precision and decrease in recall. In +Rev_new, the number of events increases around by 50%, and the model's performance significantly improves 74%, approaching optimal results. 77 competing events are eliminated in +Rev_com, leading to a increase in precision, and yet a slightly drop in recall.

In general, the key to improving the model performance lies in the extra events identified in the *Data Revision* stage.

### 6.4 Analysis on Adversarial Samples

| Setting | Models | F1 | $R_{error} \downarrow$ |
|---|---|---|---|
| zero-shot | GtR | 50.49 | **7** |
| | w/o Adv | 48.68 | 53 |
| | UIE-base | 38.14 | 63.5 |
| fully supervised | Bert-ED | **70.17** | 10.5 |

Table 4: The error rates of different models on adversarial test data, where w/o ADV indicates the model trained without adversarial samples and Bert-ED is trained on all gold samples.

We generate adversarial samples for each trigger in *Data Generation*. Previous studies have shown that incorporating adversarial samples during training can improve the robustness of models. To evaluate the effectiveness, we conduct experiments on a brand-new test set for ACE05. The detailed process for data construction is as follows:

We first select the common triggers between the gold training dataset and $\mathcal{D}_{mt}$ to ensure a fair comparison. Then, with these triggers, we create 200 adversarial samples as test set. Each sample is manually inspected to ensure that neither does it contain any events nor appear in $\mathcal{D}_{mt}$. We use the error rate $R_{error}$ as the evaluation metric, which is the proportion of incorrectly predicted samples out of all the samples.

Table 4 presents the experiments performance of different models on this test set. We find that removing the adversarial samples slightly compromises the overall performance of GtR (a 2% drop) and significantly reduces its robustness. UIE, which is not deliberately trained with adversarial samples, also exhibits high error rates. GtR trained with adversarial samples (7% $R_{error}$), and Bert-ED trained on the full training set with supervision (10.5% $R_{error}$) demonstrate much lower error rates. This finding validates the practical utility of generating adversarial data to enhance the model robustness.

### 6.5 Analysis on Weighting Model

This section analyzes the practical effects of the dynamic weighting. To do this, we first conduct a statistical analysis of the scores for each event in the ACE05 synthetic data. The horizontal axis of the graph represents the score range from 0.35 to 0.5, while the vertical axis represents the number of events corresponding to each range in the ACE05 synthetic dataset.

CCL 2024

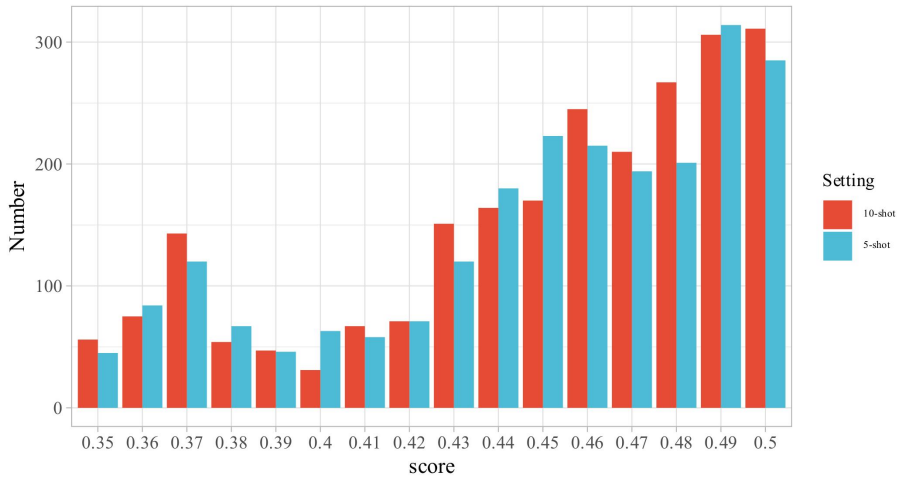



Figure 4: Illustration of the weights of events.

From the figure 4, it can be observed that the majority of synthetic data scores are concentrated in the range of 0.44 to 0.5, with some in the range of 0.35 to 0.38, and the remaining data dispersed in other score ranges. Under the 10-shot scenario, the distribution of synthetic data tends to skew slightly to the right compared to the 5-shot scenario.

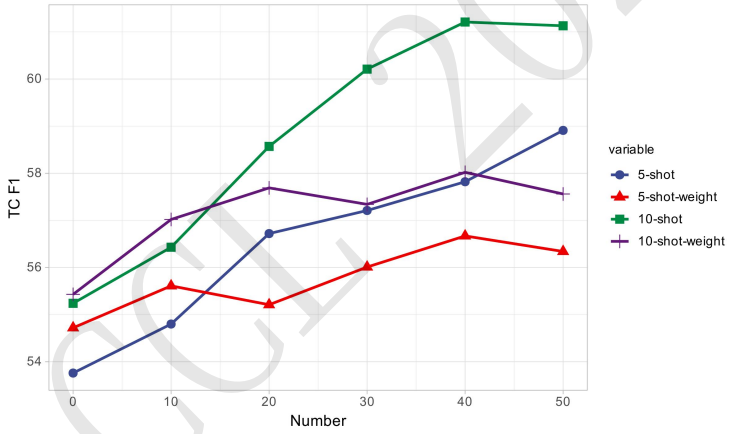



Figure 5: Illustration of the performance of models in different setting.

In Figure 5, we present a comparison of performance with and without the weighting module as the scale of synthetic data increases in both the 5-shot and 10-shot scenarios. The horizontal axis represents the number of synthetic data added for each event type, while the vertical axis represents the F1 score on TC. **5-shot-weight** and **10-shot-weight** denote the performance of models without incorporating the weighting model.

Figure 5 demonstrates that in both the 5-shot and 10-shot scenarios, models trained on dynamically weighted synthetic data outperform those trained directly on synthetic data. Under the 5-shot condition, as the data scale increases, the performance of the regular model reaches a peak, showing significant fluctuations and even slight declines. However, models trained with dynamic weighting still maintain a growth trend.

These results confirms the importance of the dynamic weighting module in improving data quality and the potential value it brings in enhancing model performance and stability.

| | Self-Bleu | Accuracy |
|---|---|---|
| Gold | 0.18 | - |
| GtR | 0.17 | 88.3 |

Table 5: Diversity and accuracy metrics of synthetic data

### 6.6 Analysis on Synthetic Data

The ablation experiment validates the role of synthetic data obtained during the data generation stage. Here, we further analyze the diversity and reliability of the synthetic data. The evaluation metric for diversity is the Self-Bleu score. We calculated the Self-Bleu scores for ACE05 gold training data and synthetic data on a randomly selected set of 300 samples. Self-Bleu values can to some extent reflect the diversity of data. It can be observed that the diversity of synthetic data is comparable to that of real data.

We manually check the accuracy of 200 adversarial data samples. Table 5 shows that the accuracy of adversarial data is relatively high, indicating that the model is proficient in synthesizing adversarial data.

### 6.7 Case Study

| Event Types | Triggers | Data Types | Sentences |
|---|---|---|---|
| Attack | bashed | normal | The reckless driver **bashed** into the cyclist, injuring him badly and causing damage to his bike. |
| Attack | seize | adversarial | The company decided to **seize** the opportunity to expand business when a rival company had to shut down due to bankruptcy |
| Be-Born | deliver | normal | The pregnant woman was rushed to the emergency room as she went into labor and had to **deliver** her premature baby. |
| Be-Born | born | adversarial | The idea for the new product was **born** during a brainstorming session with the marketing team. |

Table 6: Generated data cases on ACE05. The bold font indicates the correct events within basic data, the green words represent the newly recognized events and the red word suggest an incorrect event.

The section presents some successful and failed cases in the synthetic data, along with the results of the weight model. As shown in Table 6 if the event type is specific and easily distinguishable from other event types, GtR tends to synthesize successful and high-quality data, as exemplified by the high correctness of the synthetic data for the **Be-Born** event shown in the table. However, even after data revision, synthetic data may still contain incorrectly recognized events, such as the event in the first case where the victim of **Attack** should be a person.

Addressing the effectiveness of data synthesis on abstract event types and proposing better revision or filtering algorithms are topics worthy of future research. We leave the work for future research.

## 7 Conclusion

ED has long faced challenges of data scarcity and the time-consuming nature of manual annotation. The development of LLMs brings a ray of hope to address these issues. In this paper, we proposes GtR, the data synthesis framework, which includes three stages: trigger expansion, data generation, and data revision, along with a dynamic weighting module for synthetic data. In the experimental section, we test the generative performance of GtR in zero-shot and low-shot scenarios. Compared to direct prompting and other generation methods, our framework exhibits significant advantages, with the dynamic weighting mechanism improving the model's performance ceiling. We hope that the research in this paper can inspire future work in the field of data synthesis.

CCL 2024

## References

Xinya Du and Claire Cardie. 2020. Event Extraction by Answering (Almost) Natural Questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 671–683, Online, November. Association for Computational Linguistics.

Jun Gao, Changlong Yu, Wei Wang, Huan Zhao, and Ruifeng Xu. 2022. Mask-then-Fill: A Flexible and Effective Data Augmentation Framework for Event Extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4537–4544, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Jiahui Gao, Renjie Pi, Yong Lin, Hang Xu, Jiacheng Ye, Zhiyong Wu, Weizhong Zhang, Xiaodan Liang, Zhenguo Li, and Lingpeng Kong. 2023. Self-Guided Noise-Free Data Generation for Efficient Zero-Shot Learning, February. arXiv:2205.12679 [cs].

Ridong Han, Tao Peng, Chaohao Yang, Benyou Wang, Lu Liu, and Xiang Wan. 2023. Is Information Extraction Solved by ChatGPT? An Analysis of Performance, Evaluation Criteria, Robustness and Errors, May. arXiv:2305.14450 [cs].

I-Hung Hsu, Kuan-Hao Huang, Elizabeth Boschee, Scott Miller, Prem Natarajan, Kai-Wei Chang, and Nanyun Peng. 2022. DEGREE: A Data-Efficient Generation-Based Event Extraction Model. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1890–1908, Seattle, United States, July. Association for Computational Linguistics.

Lifu Huang, Heng Ji, Kyunghyun Cho, Ido Dagan, Sebastian Riedel, and Clare Voss. 2018. Zero-Shot Transfer Learning for Event Extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2160–2170, Melbourne, Australia, July. Association for Computational Linguistics.

Kuan-Hao Huang, I.-Hung Hsu, Tanmay Parekh, Zhiyu Xie, Zixuan Zhang, Premkumar Natarajan, Kai-Wei Chang, Nanyun Peng, and Heng Ji. 2024. TextEE: Benchmark, Reevaluation, Reflections, and Future Challenges in Event Extraction, February. arXiv:2311.09562 [cs].

Martin Josifoski, Marija Sakota, Maxime Peyrard, and Robert West. 2023. Exploiting Asymmetry for Synthetic Training Data Generation: SynthIE and the Case of Information Extraction, March. arXiv:2303.04132 [cs].

Meisin Lee, Lay-Ki Soon, Eu-Gene Siew, and Ly Fie Sugianto. 2021. An Annotated Commodity News Corpus for Event Extraction, August. arXiv:2105.08214 [cs].

Bo Li, Gexiang Fang, Yang Yang, Quansen Wang, Wei Ye, Wen Zhao, and Shikun Zhang. 2023. Evaluating ChatGPT's Information Extraction Capabilities: An Assessment of Performance, Explainability, Calibration, and Faithfulness, April. arXiv:2304.11633 [cs].

Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2022. Unified Structure Generation for Universal Information Extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5755–5772, Dublin, Ireland, May. Association for Computational Linguistics.

Mingyu Derek Ma, Xiaoxuan Wang, Po-Nien Kung, P. Jeffrey Brantingham, Nanyun Peng, and Wei Wang. 2023a. STAR: Boosting Low-Resource Event Extraction by Structure-to-Text Data Generation with Large Language Models, May. arXiv:2305.15090 [cs].

Yubo Ma, Yixin Cao, YongChing Hong, and Aixin Sun. 2023b. Large Language Model Is Not a Good Few-shot Information Extractor, but a Good Reranker for Hard Samples!, March. arXiv:2303.08559 [cs] version: 1.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is ChatGPT a General-Purpose Natural Language Processing Task Solver?, February. arXiv:2302.06476 [cs].

Timo Schick and Hinrich Schütze. 2021. Generating Datasets with Pretrained Language Models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6943–6951, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. 2019. Meta-Weight-Net: Learning an Explicit Mapping For Sample Weighting, September. arXiv:1902.07379 [cs, stat].

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. ACE 2005 Multilingual Training Corpus, February. Artwork Size: 1572864 KB Pages: 1572864 KB.

Xiaozhi Wang, Ziqi Wang, Xu Han, Wangyi Jiang, Rong Han, Zhiyuan Liu, Juanzi Li, Peng Li, Yankai Lin, and Jie Zhou. 2020. MAVEN: A Massive General Domain Event Detection Dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1652–1671, Online, November. Association for Computational Linguistics.

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022a. ZeroGen: Efficient Zero-shot Learning via Dataset Generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11653–11669, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Jiacheng Ye, Jiahui Gao, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. 2022b. ProGen: Progressive Zero-shot Dataset Generation via In-context Feedback. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3671–3683, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Hongming Zhang, Wenlin Yao, and Dong Yu. 2022a. Efficient Zero-shot Event Extraction with Context-Definition Alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7169–7179, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Senhui Zhang, Tao Ji, Wendi Ji, and Xiaoling Wang. 2022b. Zero-Shot Event Detection Based on Ordered Contrastive Learning and Prompt-Based Prediction. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2572–2580, Seattle, United States, July. Association for Computational Linguistics.

CCL 2024

# $E^3$: Optimizing Language Model Training for Translation via Enhancing Efficiency and Effectiveness

**Linqing Chen**[*1] and **Weilei Wang**[1] and **Dongyang Hu**[2]
[1]PatSnap Co., LTD., Suzhou, China
[2]Soochow University, Suzhou, China
{chenlinqing, wangweilei}@patsnap.com

## Abstract

In the field of Natural Language Processing (NLP), Large-scale Language Models (LLMs) have demonstrated exceptional capabilities across a variety of tasks, including question answering, classification, and particularly, natural language understanding. The integration of neural machine translation with LLMs presents significant potential, transforming the paradigms of cross-lingual communication and information exchange. This study investigates the foundational aspects of LLMs' translation abilities and identifies effective training methodologies to equip them with multilingual capacities. We specifically explore the optimal timing for introducing translation capabilities to LLMs via supervised tasks, considering the inherent bilingual nature of machine translation. Key questions explored include whether it is more beneficial to integrate multiple languages during the pre-training or supervised fine-tuning (SFT) stages, how variations in language ratios influence LLMs' translation abilities, and whether longer or shorter texts are more effective for training these models. This research conducts a thorough investigation by training multiple LLMs from scratch with parameter scales in the billions and enhances the robustness of our findings by upgrading the language capabilities of pre-trained open-source models with parameter scales reaching tens of billions. The aim is to provide a detailed analysis that elucidates the complexities of augmenting machine translation capabilities within LLMs.

## 1 Introduction

The Large Language Model (LLM) has proven itself capable across a broad spectrum of tasks, including named entity extraction, text classification, and document understanding. Moreover, the LLM has demonstrated the ability to address highly complex tasks using the innovative Chain of Thoughts (CoT) methodology, as pioneered by Zhang et al. (Zhang et al., 2022). Among the various capabilities of the LLM, this paper focuses on one particular area that has significantly intrigued the authors: its exceptional proficiency in machine translation. This proficiency promises transformative potential for the field, with extensive implications for cross-lingual communication and the broader dissemination of information.

The evolutionary trajectory of Large Language Models (LLMs) has undergone significant cyclical transformations, each contributing substantially to their growth and capabilities. The journey began with the introduction of the Transformer model for neural machine translation, a foundational breakthrough as described by (Vaswani et al., 2017). Building on this, BERT (Devlin et al., 2019) leveraged the Transformer's encoder to achieve exceptional performance in text classification and cloze tests.

Following the advent of BERT, a series of influential models emerged, notably GPT (Radford et al., 2019), which is characterized by its unique decoder-only architecture. Furthermore, models like T5 (Raffel et al., 2019) and mT5 (Xue et al., 2020), with their comprehensive encoder-decoder frameworks, have continuously expanded the boundaries of what is possible with LLMs. Most of these models are either directly based on the Transformer architecture or represent substantial enhancements to it, reaffirming the critical role of the Transformer in the advancement of the field. This progression of models, each aug-

*Corresponding author.
©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1023-1034, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

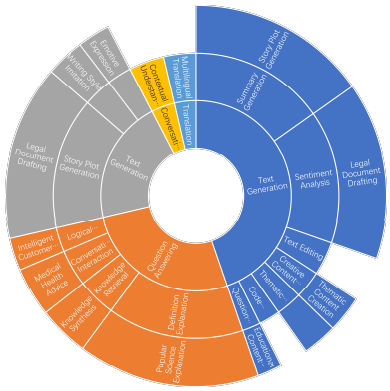



Figure 1: Capabilities of large scale language model.

menting the work of its predecessors, has driven LLMs to their current state of remarkable performance and vast potential in natural language processing (NLP).

This paper aims to delve deeper into these evolving aspects of natural language processing, particularly focusing on the translation proficiency of Pre-trained Language Models (PLMs). Our objectives are threefold: to explore the foundational aspects of translation efficiency in PLMs, to investigate the impact of diverse training methodologies and data ratios on PLMs' translation capabilities, and to propose effective strategies for training PLMs to enhance their translation performance. We explore several pertinent questions in the context of neural machine translation using LLMs:

- **Multiple Languages Alignment**: Should the alignment of language semantic spaces be performed during the pre-training stage or the supervised fine-tuning (SFT) stage?

- **Efficient Training**: Which type of contextual input—long or short—enhances the training efficiency and effectiveness of Large-scale Language Models for translation tasks?

- **Language Proportion Impact**: How does the composition of different languages in the corpus during LLMs' pre-training process affect their performance in translating various languages?

## 2 Machine Translation in the Large Language Model (LLM) Era

The exploration of machine translation driven by Language Models (LMs) has been revitalized by the advent of Large Language Models (LLMs), building on a rich history of significant research. Pioneering studies such as Brants et al. (2007) (Brants et al., 2007) integrated large-scale statistical language modeling into machine translation, laying the groundwork for subsequent advancements. This approach was refined by Mohit et al. (2009) (Mohit et al., 2009), who tailored LMs to improve the translation of complex phrases. The shift to neural language models, notably researched by Vaswani et al. (2013) (Vaswani et al., 2013), emphasized their potential to enhance translation accuracy and fluency. Lee (2020) (Lee, 2020) effectively bridged theoretical research with practical applications, particularly in quality estimation tasks, demonstrating the real-world impact of these developments.

In more recent studies, Han et al. (2022) (Han et al., 2022) examined the superiority of extra-large pre-trained language models (xLPLMs) over smaller counterparts in domain-specific translation tasks, providing vital insights into model selection for specific contexts. Concurrently, Tan et al. (2022) (Tan et al., 2022) introduced Multi-Stage Prompting, an innovative method that utilizes pre-trained LMs for translation, illustrating the continued evolution in LM applications in translation tasks. Zhu (2023) (Zhu et al., 2023) extended this investigation, exploring the capability of LLMs in translating a diverse array of languages and thus contributing to our understanding of the scalability and adaptability of LLMs in multilingual translation scenarios.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1023-1034, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

## 3 Experimentation

We explore and analyze the performance of translation of multiple Large Language Models (LLMs) from scratch, with parameter scales reaching the order of several billion. Simultaneously, given that training models with hundreds of billions of parameters incurs costs in the order of millions of dollars, when investigating questions on an even larger scale of Large Language Models (LLMs), we opt to leverage open-source LLMs.

### 3.1 Training LLM from Scratch

The development of expansive Large Language Models (LLMs) is a resource-intensive endeavor, requiring substantial investments in both time and financial resources. Previous research (Henighan et al., 2020) has systematically demonstrated that insights garnered from experiments on smaller-scale LLMs can reliably inform the development of their larger-scale versions. This is substantiated by the rigorous empirical investigations conducted by leading organizations such as OpenAI (OpenAI, 2023) and Google (Longpre et al., 2023). These studies primarily focus on Pretrained Language Models (PLMs) with parameters numbering in the billions, which, although substantial, are comparatively modest against the backdrop of the most extensive LLMs. Such research underscores the practical relevance and extrapolative potential of smaller LLM findings to larger model architectures, thereby justifying continued investment in their development.

Opting for models of a more manageable scale confers distinct advantages, notably in accelerating the pace of model iteration. This acceleration facilitates a more efficient exploration of the research landscape in natural language processing. Furthermore, this strategy provides us with a valuable opportunity to engage in the crucial task of contrasting and evaluating diverse phenomena within the field. Such activities are essential for broadening the collective understanding and advancing the state of knowledge in this rapidly evolving domain.

#### 3.1.1 Training Data

Detailed information regarding the datasets used is thoroughly documented in Table 1.

| Stage | Data | Size(GB) | Line(M) |
|---|---|---|---|
| Pre-Training | Wikipedia EN | 20.0 | 8.5 |
| | Wikipedia ZH | 2.4 | 2.0 |
| | news2016zh | 7.5 | 1.3 |
| Translation | translation2019 | 1.6 | 5.1 |
| | LDC zh-en news | 0.3 | 0.8 |

Table 1: Training Data.

**Pre-Training Data** For the initial pre-training of our Pre-trained Language Models (PLMs), we utilize both the Chinese Wikipedia[0] and the English Wikipedia[1] as the foundational corpora. Due to the relatively limited size of the Chinese Wikipedia dataset, we enhance our Chinese pre-training corpus with additional datasets, specifically the news2016zh (brightmart, 2019) and the translation2019zh[2], to enrich the training data significantly. This supplementation of the Chinese corpus is crucial for maintaining a proportional balance in the volume of language data between Chinese and English, particularly in ablation studies designed to evaluate the impact of dataset size on model performance.

**Fine-Tuning Data** To achieve Chinese-to-English (ZH-EN) translation capabilities in our models, we have selected the Linguistic Data Consortium (LDC) corpora as our training data for translation tasks. We chose this corpus due to its unique properties; it not only serves as a resource for sentence alignment but also for document alignment. Inspired by Zhang et al. (Zhang et al., 2018), we utilize different forms

[0] https://dumps.wikimedia.org/zhwiki/
[1] https://dumps.wikimedia.org/enwiki/
[2] https://github.com/brightmart/nlp_chinese_corpus

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1023-1034, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

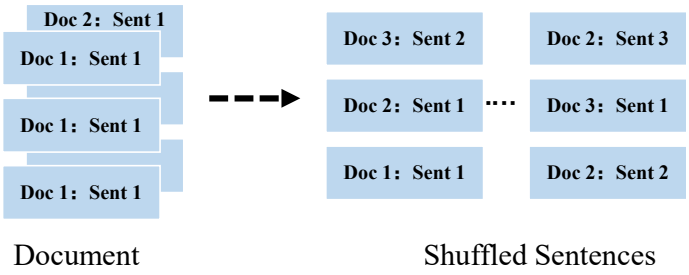



Figure 2: In the context of translation training data, retaining the inherent document structure is termed as document-level translation data. Conversely, when parallel sentence pairs within such data are isolated from their document context, this process results in what is known as sentence-level parallel data.

of this parallel corpus under various experimental setups. When evaluating the effectiveness of short-text or context-free translation data, we employ the sentence-aligned version of the corpus. Conversely, for assessing the performance with long-text or rich-context translation data, we use the document-aligned version of the corpus. Both forms are consistent in content and volume, but differ in their alignment, which is illustrated in Figure 2. When using the sentence-aligned version, we segment the originally document-aligned corpus into sentences.

### 3.1.2 Pre-Training

| Hyper Parameters | Pre-training | Finetuning |
|---|---|---|
| Model | 1.3B | 1.3B |
| mini batch size | 16 | 8 |
| global batch size | 128 | 64 |
| Min_LR | 1e-5 | 1e-6 |
| Max_LR | 3e-5 | 1e-5 |
| Max_len | 2048 | 4096 |
| TP | 8 | 8 |
| PP | 8 | 8 |

Table 2: Details of training settings.

We have adopted an architectural framework akin to that of LLaMA (Touvron et al., 2023a) and have undertaken the training of our model from scratch. The AdamW optimizer (Loshchilov and Hutter, 2018) is employed for training, with hyperparameters $\beta_1$ and $\beta_2$ set to 0.9 and 0.95, respectively. We apply a weight decay of 0.1 and enforce a gradient norm clipping of 0.5. Unless otherwise stated, the base model in this subsection is configured with 1.3 billion parameters. The maximum sequence lengths are set to 2048 tokens for pre-training and 4096 tokens for fine-tuning. In the pre-training phase, when supervised tasks like machine translation are incorporated, the computation of the input's loss is excluded to prevent the model from predicting sentences in the source language during training. Machine translation evaluations in this paper are reported using the BLEU score (Papineni et al., 2002), unless specified otherwise. On average, training each model requires one week on 8 A800 GPUs. Further details regarding training settings are presented in Table 2.

### 3.1.3 Fine-Tuning

In the experiments conducted in this subsection, we employ Supervised Fine-Tuning (SFT) to endow the model with translation capabilities. The SFT used here differs slightly from that in other research works. On one hand, we require SFT to enable the model to understand the task of mapping the source language to the target language. On the other hand, we do not need a diverse set of SFT tasks since capabilities beyond translation are not required. For the same reasons, we also forgo the Reinforcement Learning from Human Feedback (RLHF) process.

CCL 2024

| Method | No. | CTX | MT02 | MT03 | MT04 | MT05 | MT08 | Average |
|---|---|---|---|---|---|---|---|---|
| NTP+SUP | #1 | Non-Concat | 42.39 | 42.30 | 41.99 | 42.35 | 31.88 | 40.07 |
| | #2 | Concat | 43.10 | 43.27 | 42.33 | 42.59 | 32.20 | 40.98 |
| NTP+MIX | #3 | Non-Concat | 43.84 | 44.10 | 41.93 | 42.89 | 32.90 | 40.81 |
| | #4 | Concat | 44.95 | 44.80 | **44.63** | **44.71** | **36.50** | **43.01** |
| Holistic MIX | #5 | Non-Concat | 43.55 | 44.27 | 41.85 | 42.90 | 32.64 | 41.04 |
| | #6 | Concat | **44.01** | **44.89** | 44.57 | 44.61 | 36.03 | 42.82 |

Table 3: NTP is an abbreviation for 'next token prediction,' and SUP stands for 'supervised task training'. 'Concat' and 'Non-concat' represent whether to concatenate supervised training data.

#### 3.1.4 Experimental Results

**Languages Alignment Stage** At what stage does aligning different languages enhance the multilingual capabilities of models? Contemporary state-of-the-art Large Language Models (LLMs) such as GPT (Radford et al., 2019) and LLaMA (Touvron et al., 2023a) typically do not utilize parallel corpora during the pre-training phase. To explore the effect of language alignment on the translation proficiency of LLMs, we initiated an investigation involving the training of two distinct types of base models. Table 3 lists all experimental results about language alignment in different stages. NTP is an abbreviation for 'next token prediction,' and SUP stands for 'supervised task training.'

- **NTP+SUP**, represents a two-step training process where pure token prediction pretraining is performed first, followed by supervised translation task training, with both training phases kept entirely separate.
- **NTP+MIX**, denotes a two-step training approach where pure token prediction pretraining is conducted initially, followed by mixed supervised and unsupervised task training.
- **Holistic MIX**, refers to a unified training approach that combines pure token prediction pre-training with translation task data in a completely integrated manner.
- **Concat**, denotes the method of concatenating supervised training data. In the 'Concat' configuration, supervised translation data is concatenated to fully utilize tokens within sequences of up to 2048 in length during the pre-training phase.
- **Non-Concat**, setting does not concatenate the data, with each sequence containing only one source-target sentence pair.

The experimental results presented in Table 3 clearly demonstrate that Model #3 achieves superior translation performance compared to Model #1 and Model #5, while Model #4 outperforms Model #2 and Model #6 in terms of translation quality. Keeping the volume of data constant, it is apparent that segregating supervised and self-supervised training phases does not contribute positively to the translation efficacy of Large Language Models (LLMs). In contrast, integrating self-supervised and supervised training approaches—particularly after completing the self-supervised next token prediction task—significantly improves translation performance. This improvement suggests that mixed training methodologies may enable models to concurrently develop foundational background knowledge and enhance their capabilities in aligning semantic spaces of two different languages during translation task training.

**Long or Short?** Further analysis of the experimental results in Table 3 indicates that Model #2 exhibits superior translation performance compared to Model #1, while Model #4 outshines Model #3 in terms of translation quality. With the data volume held constant, it is found that concatenating supervised translation data significantly improves the translation capabilities of Large Language Models (LLMs). These findings suggest that concatenated supervised translation training data not only optimizes the use of sequence length, thereby reducing the number of data samples needed, but also allows for a higher

CCL 2024

number of sentence pairs to be processed simultaneously. This efficient approach to data utilization evidently enhances the translation performance of LLMs.

A closer examination of the experimental results unveils an intriguing phenomenon. Despite being based on differing training methodologies, Model #3 exhibits only marginal improvements in translation performance over Model #2 and, in some cases within individual sub-test sets, performs worse. This observation underscores the importance of fully utilizing sequence lengths in training data and ensuring data richness within these sequences to effectively boost the translation performance of Large Language Models (LLMs). Consequently, for the training and development of LLMs, the employment of apt data utilization strategies and alignment methods is critical for enhancing translation capabilities.

**Translation Performance** It is worth noting that the authors of this paper have observed that the performance of simply trained PLMs on translation tasks does not surpass that of dedicated translation models. This finding exceeds initial expectations and aligns with the research conducted by (Zhu et al., 2023). In both concatenation and non-concatenation translation tasks, PLMs, functioning solely as decoders, have not achieved superior performance compared to specialized translation models. We speculate that this limitation could be attributed to the decoder-only architecture of PLMs. However, it is conceivable that with further increases in model parameters and training data or the adoption of models that incorporate both encoder and decoder components, such as mT5, PLMs may eventually surpass dedicated translation models.

| Model | MT02 | MT03 | MT04 | MT05 | MT08 | BLEU |
|---|---|---|---|---|---|---|
| #1 ZH-EN 2:1 | 37.09 | 36.92 | 36.54 | 34.68 | 27.74 | 34.89 |
| #2 ZH-EN 1:1 | 36.58 | 36.40 | 35.87 | 36.23 | 30.08 | 34.05 |
| #3 ZH-EN 1:2 | 37.17 | 36.85 | 36.31 | 34.61 | 27.72 | 34.50 |

Table 4: The impact of baseline models trained with different language ratios on downstream translation tasks is investigated. ZH-EN x:y denotes baseline models with a Chinese-English language ratio of x:y.

**Language Ratio** We employed different ratios of Chinese and English Wikipedia data to investigate the impact of non-translation task data proportion during the pretraining stage on the performance of downstream translation tasks. The model named "EN-ZH 1:2" represents a base model where English Wikipedia data accounts for 1/3, and Chinese data accounts for 2/3 of the total. The proportions of the other two models can be inferred from their names accordingly. During the supervised fine-tuning stage, we introduced LCD zh-en news translation tasks to assess the influence of different language proportions during the pretraining stage on the performance of downstream translation tasks.

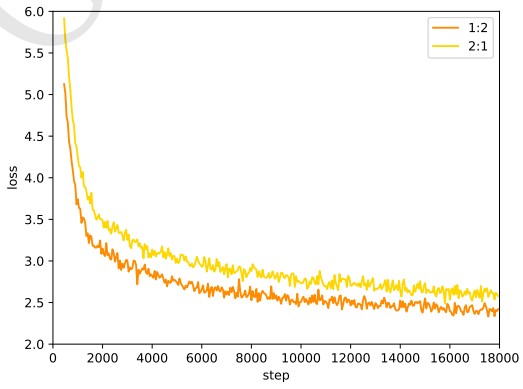



Figure 3: Illustration of two different models('ZH-EN 2:1 and 'ZH-EN 1:2') loss.

Figure 3 illustrates the loss curves of two base models with Chinese-English data ratios of 1:2 and 2:1 during the training process. It can be observed that the base model with a higher proportion of Chinese

data converges more slowly. Initially, the authors of this paper believed that the slower loss reduction observed in the base model with a Chinese-to-English ratio of 2:1 might result in relatively poorer translation performance, because the Chinese language is harder to learning for LLM. The experimental data in Table 4 indicates that, in terms of zh-en translation tasks, while different language ratios may affect the convergence speed or rate of loss reduction of the base models, **there is no significant difference observed in downstream translation tasks**.

### 3.2 Post-Training on Open-Source LLM

In the preceding sections, we engaged with a range of intriguing questions and observed phenomena based on several models, each possessing 1.3 billion parameters, trained from scratch. This naturally leads us to a critical inquiry: Are these phenomena universal, and do they persist across larger-scale LLMs? To address these questions within reasonable time and budget constraints, we extended the Chinese language capabilities of the open-source 13 billion parameter model and reaffirmed the phenomena previously observed in the 1.3 billion parameter model.

#### 3.2.1 Training Data

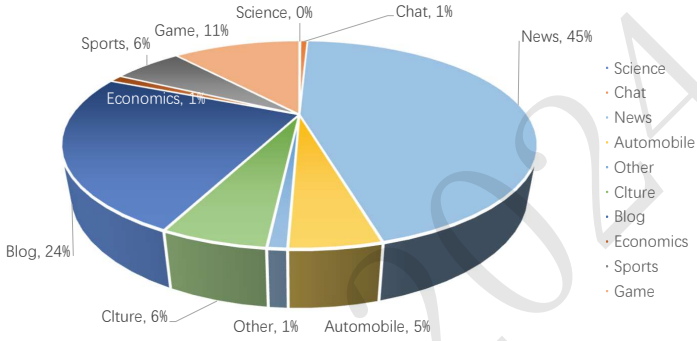



Figure 4: The composition of training data types employed to enhance the Chinese language capabilities of the model.

We examine the impact of various training methods on the translation performance by focusing on enhancing the English-to-Chinese translation capabilities of our base models. To this end, we have randomly sampled 40 billion tokens from the Wudao corpus (Yuan et al., 2021), aiming to improve the English-to-Chinese translation proficiency. The distribution of data types from this corpus is depicted in Figure 4.

#### 3.2.2 Post-Training

We have trained the model using the 13 billion parameter version of LLaMA2 (Touvron et al., 2023b). Unless otherwise noted, the maximum sequence length for data is set at 4096 tokens. During the pre-training phase, when supervised tasks such as machine translation are integrated, the computation of the input's loss is omitted to prevent the model from merely reproducing sentences from the source language. Unless specified differently, the evaluation of machine translation performance in this study is based on the BLEU score. For the training of these models, an average duration of 10 days on 64 A800 GPUs per model is required.

#### 3.2.3 Fine-Tuning

Similar to Section 3.1, the models in this subsection, which continue training on LLaMA2, utilize Supervised Fine-Tuning (SFT) to enable the model to comprehend the machine translation task. This approach specifically suppresses other capabilities of the model to prevent interference with the automated evaluation of its translation performance.

CCL 2024

| Model | Concat | MT02 | MT03 | MT04 | MT05 | MT08 | Average |
|---|---|---|---|---|---|---|---|
| #1 | Holistic Mix | 48.88 | 48.20 | 44.58 | 48.01 | 40.97 | 46.33 |
| #2 | NTP + Sup | 47.90 | 47.30 | 43.88 | 47.20 | 40.08 | 45.07 |
| #3 | NTP + Mix | **49.59** | **48.91** | **45.22** | **48.69** | **41.52** | **47.00** |
| Model | Non-Concat | MT02 | MT03 | MT04 | MT05 | MT08 | Average |
| #4 | Holistic Mix | 46.41 | 46.24 | 44.53 | 46.44 | 40.50 | 45.37 |
| #5 | NTP + Sup | 45.59 | 46.21 | 45.47 | 45.43 | 35.00 | 44.01 |
| #6 | NTP + Mix | 46.89 | 46.64 | 44.90 | 46.89 | 40.98 | 45.80 |

Table 5: NTP is an abbreviation for 'next token prediction,' and SUP stands for 'supervised task training.'. 'Concat' and 'Non-concat' represent whether to concatenate supervised training data.

#### 3.2.4 Experimental Results

In our research, we aimed to enhance the Chinese language capabilities of open-source Large Language Models (LLMs) by adhering to established experimental protocols similar to those used for training models from scratch. The training of translation tasks was segmented into distinct phases, and a comparative analysis of the outcomes was conducted.

**Languages Alignment Stage** We investigated the optimal timing for imparting translation capabilities to LLMs through supervised tasks. According to the results presented in Table 5, Model #1 showed superior translation performance compared to Model #2, and similarly, Model #4 outperformed Model #5. These training experiments, designed to enhance translation capabilities in open-source LLMs, are consistent with our previous findings. Notably, maintaining a constant volume of training data and integrating the next token prediction (NTP) task with supervised translation training has proven effective in enhancing the translation performance of LLMs. Additionally, our experiments revealed that LLMs, particularly those with billions of parameters, can sometimes misinterpret translation tasks when trained on large and diverse datasets, leading to undesired outputs. To address this, we maintained a constant volume of training data but extracted a small subset for translation training. Following the MIX training phase, we conducted targeted machine translation training to refine the model's translation capabilities. This focused approach resulted in improved translation outcomes.

**Long or short?** How to train the translation capability of Large-scale Language Models more efficiently. The experiments in the Table 5 are broadly categorized into two main groups: concatenating or not concatenating the translation data. It is evident that, on the whole, the translation approach using concatenated training data outperforms the non-concatenated approach. Furthermore, based on our calculations, the time consumed for training with concatenated data is slightly lower, further substantiating its efficiency.

### 3.3 Ablation Experiments

In this section, we explore additional critical factors influencing the training of Large Language Models (LLMs) with translation capabilities. Specifically, we examine the effects of the length of translation pairs during the training stage and the model's performance in domains requiring specialized knowledge.

#### 3.3.1 Length of Training Language Paris

We utilize context-aware BLEU (d-BLEU) to assess the impact of text length on the translation performance of Large Language Models (LLMs). The ablation experiments in this section build upon the models from the language ratio impact experiments detailed in Section 3.1.4. These models were trained using varying language ratios in sentence pairs to examine their translation capabilities. Although the maximum length of training sequences is set to 4096, many tokens are often filled with the special token [pad], denoting **None**, due to the varying lengths of sentences.

**Context Aware Template** In the ctx-awar experiments, each sentence is precisely designated as the target for translation. Concurrently, the sentences preceding and following the target sentence are seam-

CCL 2024

lessly concatenated to provide the necessary contextual backdrop. To construct the input, we adeptly utilize a structured template: “Translate [sentence] based on the contextual foundation: [context].”

| Ratio | No. | CTX | MT02 | MT03 | MT04 | MT05 | MT08 | BLEU |
|---|---|---|---|---|---|---|---|---|
| ZH-EN 2:1 | #1 | + sent-level | 37.09 | 36.92 | 36.54 | 34.68 | 27.74 | 34.89 |
| | #2 | + ctx-aware | 36.52 | 35.85 | 35.02 | 35.79 | 27.10 | 34.44 |
| ZH-EN 1:1 | #3 | + sent-level | 36.58 | 36.40 | 35.87 | 36.23 | 30.08 | 34.05 |
| | #4 | + ctx-aware | 36.80 | 36.55 | 36.28 | 36.57 | 29.65 | 34.33 |
| ZH-EN 1:2 | #5 | + sent-level | 37.17 | 36.85 | 36.31 | 34.61 | 27.72 | 34.50 |
| | #6 | + ctx-aware | 37.01 | 36.84 | 36.52 | 36.77 | 27.15 | 34.86 |

Table 6: The impact of baseline models trained with different language ratios on downstream translation tasks is investigated. ZH-EN x:y denotes baseline models with a Chinese-English language ratio of x:y, ’+ sent-level’ refer to using sentence-level ZH-EN news data as the downstream translation task. ”ctx-aware” refers to downstream tasks using the same data with context-aware template, with details provided in the Section 3.3.1.

**Results** As indicated in Table 6, Models #2, #4, and #6 perform context-aware translation tasks on Models #1, #3, and #5, respectively, using context-aware templates. These templates are utilized to assess each model’s capability with long sequences. Our hypothesis is that if training with short pairs does not negatively impact the model’s translation performance, then the BLEU scores for context-aware translation tasks should be approximately 1 point higher than those for sentence-level tasks, due to the provision of contextual information. The experimental results, as shown in Table 3 Section 3.1, suggest a similar trend. The performance of Models #6, #4, and #2 is only marginally better than that of Models #5, #3, and #1. This indicates that limited exposure to shorter text sequences during the sentence-level translation training in Models #1, #3, and #5 has adversely impacted their ability to process longer texts, resulting in a decline in performance on context-aware tasks with lengthier sequences.

### 3.3.2 Translation in the Biomedical Domain

| Model | Concat | ZH-EN | EN-ZH |
|---|---|---|---|
| #1 | Holistic Mix | 18.60 | 20.11 |
| #2 | NTP + Sup | 18.19 | 19.12 |
| #3 | NTP + Mix | 18.98 | 20.25 |
| Model | Non-concat | MT02 | MT03 |
| #4 | Holistic Mix | 17.58 | 19.25 |
| #5 | NTP + Sup | 17.33 | 19.10 |
| #6 | NTP + Mix | 17.90 | 19.91 |

Table 7: NTP is an abbreviation for ’next token prediction,’ and SUP stands for ’supervised task training.’. ’Concat’ and ’Non-concat’ represent whether to concatenate supervised training data.

To validate the results obtained in earlier chapters from various perspectives, we conducted zero-shot translation using publicly available translation test sets that encompass extensive domain knowledge. Zero-shot machine translation in this context refers to an experiment where, unlike previous experiments, we did not first enhance the base model’s Chinese capabilities using next token prediction with Chinese data. Instead, we directly trained on very limited translation corpora using different training methodologies for domain-specific translation tasks. We performed bidirectional zero-shot experiments on the publicly available translation test set in the biomedical field (Bawden et al., 2020). The results, as shown in Table 7, are consistent with the early findings using models trained as described in Table 5.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1023-1034, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

## 4 Related Work

In the study by (Brants et al., 2007), a comprehensive examination is provided regarding the advantages stemming from the integration of large-scale statistical language modeling within the domain of machine translation. Additionally, (Mohit et al., 2009) conduct a thorough investigation into the intriguing notion of adapting language models specifically tailored for phrases exhibiting suboptimal translation quality. The research conducted by (Vaswani et al., 2013) delves into the intricate realm of neural language models and their potential application within the domain of machine translation, shedding light on the innovative techniques that harness the power of these models. In a detailed analysis, (Lee, 2020) elucidate the contributions made by Bering Lab to the WMT 2020 Shared Task on Quality Estimation (QE), underscoring the significance of their submission in advancing the field.

In a quest to explore the nuances of performance differences, (Han et al., 2022) embark on a meticulous examination aimed at ascertaining whether extra-large pre-trained language models (xLPLMs) unequivocally surpass their smaller-sized counterparts for domain-specific machine translation tasks. Expanding the horizon of possibilities, (Tan et al., 2022) introduce the concept of Multi-Stage Prompting (MSP), a straightforward yet impactful approach designed to harness the capabilities of pre-trained language models for translation tasks. Delving into the intricacies of multilingual translation, (Zhu et al., 2023) undertake an in-depth exploration of the performance characteristics exhibited by large language models (LLMs), simultaneously delving into the various factors that exert influence on the translation prowess of these models.

## 5 Conclusion

In this study, we conducted an extensive series of comparative experiments using five models trained from scratch, each with a parameter count of 1.3 billion, alongside several continued-training scenarios on open-source Large Language Models (LLMs) of 13 billion parameters. Our key findings reveal that integrating translation tasks during the pre-training phase of LLMs consistently enhances performance in handling longer text sequences, especially evident in context-aware translation tasks. Additionally, our results indicate that strategically segmenting translation training tasks from the base model training into distinct phases can significantly expedite the overall training process of the model. While variations in language ratios were observed to affect the convergence speed and training duration of the base model, they did not substantially impact the performance metrics of downstream translation tasks. This dynamic interaction between pre-training strategies, task integration, and language ratio effects underscores the potential for optimizing both the efficiency and effectiveness of neural machine translation systems within large-scale language models. Future efforts will focus on employing under-resourced languages to investigate methods for bridging 'Zero-Shot' language translation challenges.

CCL 2024

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1023-1034, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

## References

Rachel Bawden, Giorgio Maria Di Nunzio, Cristian Grozea, Inigo Jauregi Unanue, Antonio Jimeno Yepes, Nancy Mah, David Martinez, Aurélie Névéol, Mariana Neves, Maite Oronoz, Olatz Perez-de Viñaspre, Massimo Piccardi, Roland Roller, Amy Siu, Philippe Thomas, Federica Vezzani, Maika Vicente Navarro, Dina Wiemann, and Lana Yeganova. 2020. Findings of the WMT 2020 biomedical translation shared task: Basque, Italian and Russian as new additional languages. In *Proceedings of the Fifth Conference on Machine Translation*, pages 660–687, Online, November. Association for Computational Linguistics.

Thorsten Brants, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean. 2007. Large language models in machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 858–867, Prague, Czech Republic, June. Association for Computational Linguistics.

brightmart. 2019. brightmart/nlp_chinese_corpus: release version 1.0, September.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805.

Lifeng Han, Gleb Erofeev, Irina Sorokina, Serge Gladkoff, and Goran Nenadic. 2022. Examining large pre-trained language models for machine translation: What you don't know about it. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 908–919, Abu Dhabi, United Arab Emirates (Hybrid), December. Association for Computational Linguistics.

T. J. Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B. Brown, Prafulla Dhariwal, Scott Gray, Chris Hallacy, Benjamin Mann, Alec Radford, Aditya Ramesh, Nick Ryder, Daniel M. Ziegler, John Schulman, Dario Amodei, and Sam McCandlish. 2020. Scaling laws for autoregressive generative modeling. *ArXiv*, abs/2010.14701.

Dongjun Lee. 2020. Two-phase cross-lingual language model fine-tuning for machine translation quality estimation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1024–1028, Online, November. Association for Computational Linguistics.

S. Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David M. Mimno, and Daphne Ippolito. 2023. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. *ArXiv*, abs/2305.13169.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Behrang Mohit, Frank Liberato, and Rebecca Hwa. 2009. Language model adaptation for difficult to translate phrases. In Lluís Màrquez and Harold L. Somers, editors, *Proceedings of the 13th Annual conference of the European Association for Machine Translation, EAMT 2009, Barcelona, Spain, Map 14-15, 2009*. European Association for Machine Translation.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Kishore Papineni, Salim Roukos, Ward Todd, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *ArXiv*, abs/1910.10683.

Zhixing Tan, Xiangwen Zhang, Shuo Wang, and Yang Liu. 2022. MSP: Multi-stage prompting for making pre-trained language models better translators. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6131–6142, Dublin, Ireland, May. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aur'elien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

CCL 2024

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Yinggong Zhao, Victoria Fossum, and David Chiang. 2013. Decoding with large-scale neural language models improves translation. In *Conference on Empirical Methods in Natural Language Processing*.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. In *North American Chapter of the Association for Computational Linguistics*.

Sha Yuan, Hanyu Zhao, Zhengxiao Du, Ming Ding, and Jie Tang. 2021. Wudaocorpora: A super large-scale chinese corpora for pre-training language models. *AI Open*.

Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, Min Zhang, and Yang Liu. 2018. Improving the transformer translation model with document-level context. In *Conference on Empirical Methods in Natural Language Processing*.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Jiajun Chen, Lei Li, and Shujian Huang. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *ArXiv*, abs/2304.04675.

CCL 2024

# Multi-features Enhanced Multi-task Learning for Vietnamese Treebank Conversion

**Zhenguo Zhang**[1,2], **Jianjian Liu**[1,2], **Ying Li**[1,2*]
[1]Yunnan Key Laboratory of Artificial Intelligence, Kunming, China
[2]Faculty of Information Engineering and Automation,
Kunming University of Science and Technology, Kunming, China
zzg@stu.kust.edu.cn,1807164254@qq.com,yingli_hlt@foxmail.com

## Abstract

Pre-trained language representation-based dependency parsing models have achieved obvious improvements in rich-resource languages. However, these model performances depend on the quality and scale of training data significantly. Compared with Chinese and English, the scale of Vietnamese Dependency treebank is scarcity. Considering human annotation is labor-intensive and time-consuming, we propose a multi-features enhanced multi-task learning framework to convert all heterogeneous Vietnamese Treebanks to a unified one. On the one hand, we exploit Tree BiLSTM and pattern embedding to extract global and local dependency tree features from the source Treebank. On the other hand, we propose to integrate these features into a multi-task learning framework to use the source dependency parsing to assist the conversion processing. Experiments on the benchmark datasets show that our proposed model can effectively convert heterogeneous treebanks, thus further improving the Vietnamese dependency parsing accuracy by about 7.12 points in LAS.

## 1 Introduction

Dependency parsing, as a fundamental task of natural language processing, aims at capturing syntactic and semantic information between two words through a dependency tree. The top of Figure 1 shows a dependency tree from the universal dependency treebank, where the directed edge from the headword "mừng (happy)" to the modified word "mà (and)" is a dependency arc, the label "mark" on the arc indicates the type of dependency relation between two words. Recently, dependency parsing has gained more attention due to its simple and understandable structure. The results of dependency parsing have been applied to various artificial intelligence tasks, such as grammatical error correction (Li et al., 2022), machine translation (Pu and Sima'an, 2022), and semantic role labeling (Zhang et al., 2022).

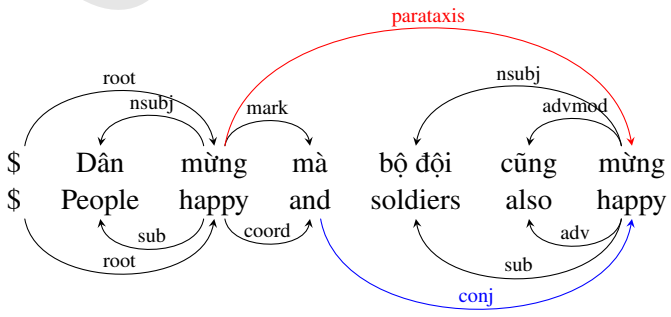



Figure 1: Example of treebank conversion from the target-side UD tree (upper) to the source-side VnDT tree (under).

©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1035-1046, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

With the development of neural networks and pre-trained language technologies, dependency parsing models have achieved significant progress in rich-resource languages. Currently, the supervised dependency parsing model has obtained 96.4 LAS and 97.4 UAS on English Penn Treebank (Amini et al., 2023), and 92.5 LAS and 93.5 UAS on Chinese Penn Treebank (Yang and Tu, 2022). However, all these supervised parsing models are highly dependent on the quality and scale of training data. For low-resource languages like Vietnamese, the performance and efficiency of the parsing model are not ideal because of the scarcity of Vietnamese annotation corpora. To address this issue, we first survey the existing Vietnamese dependency treebanks, which include Universal Dependencies (UD)[0], Vietnamese Dependency Treebank (VnDT) (Nguyen et al., 2014a), and Vietnamese Language and Speech Processing (VLSP2020)[1]. However, since these treebanks adopt different annotation standards, it is impossible to fuse all treebanks directly for traditional paring model training due to the potential semantic conflicts. As shown in Figure 1, the upper tree is generated based on the UD standard, while the bottom tree is from VnDT. We can see that even though the two dependency trees have the same input sentence, the tree structures still have some differences in arcs and labels. This study terms such treebanks with different standards or linguistic theories as heterogeneous treebanks. Hence, how to convert the heterogeneous treebanks to a unified one is the main challenge to improving the performance of Vietnamese dependency parsing. The key to addressing this problem is to model the mappings between different treebanks. In the past few years, some researchers have attempted to perform conversion between heterogeneous treebanks through rule-based methods (Stymne et al., 2018; Borges Völker et al., 2019; Arnardóttir et al., 2020). However, most of these conversions focus on bi-tree alignment, ignoring the information in the source dependency tree. Hence, we propose a multi-features enhanced multi-task learning method to convert other treebanks into the UD standard. Concretely, we first use the pre-trained language model XLM-RoBARTa to enhance the basic dependency parsing performance. Then, we exploit pattern embedding and Tree BiLSTM to extract the local and global features from the source treebank. Next, both features are fused into a multi-task learning framework to further enhance the capability of conversion. Finally, all converted trees are combined with the original UD for the XLM-RoBARTa-enhanced basic model training.

Experimental results demonstrate that our proposed model can effectively model the mapping between heterogeneous treebanks, and the converted treebank is extremely useful for improving the accuracy of Vietnamese dependency parsing. Detailed analysis shows that both pattern embedding and Tree BiLSTM are helpful in encoding the information of the source tree, thus improving the conversion performance. The comparison experiments indicate that the converted treebank is an important factor in boosting the Vietnamese dependency parsing.

## 2 Related work

### 2.1 Dependency Parsing

In the field of dependency parsing, the prevailing methods are graph-based and transition-based dependency parsing. Graph-based dependency parsing has achieved state-of-the-art performance in various languages. With the development of neural networks and pre-trained language models, an increasing number of researchers and developers are applying these technologies to dependency parsing models. These models have a strong capability to understand complex linguistic structures, achieving significant progress in dependency parsing tasks. Dozat and Manning (2017) proposed a biaffine parser model, which uses a BiLSTM for encoder layer. Li et al. (2018) proposed the first end-to-end model for dependency parsing. Kondratyuk and Straka (2019) developed a multilingual multi-task self-attention network fine-tuned on BERT, achieving competitive results in multiple languages. Li et al. (2020) employed a bi-directional LSTM-CNN architecture for context encoding and deep contextually related word representations (ELMo and BERT) to further enhance model performance. Sharefah et al. (2023) delved into fine-tuning the BERT model for Arabic dependency parsing, providing insights for fine-tuning BERT models for specific NLP tasks.

[0] https://universaldependencies.org/
[1] https://vlsp.org.vn/vlsp2020/eval/udp

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1035-1046, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Vietnamese is part of the Austro-Asiatic language family and is one of the 20 most commonly used languages worldwide. Vietnamese vocabulary contains a large number of Sino-Vietnamese words derived from Chinese. This unique language structure and vocabulary combination makes Vietnamese an interesting research object in the field of NLP. Especially in the field of Vietnamese dependency parsing, significant research progress has been made in recent years. Do et al. (2022) proposed a cross-linguistic model adaptation using English to aid Vietnamese parsing, enhancing accuracy. Thien et al. (2020) combined BiLSTM and GNN. Le-Hong and Cambria (2024) integrated graph embedding with bidirectional recurrent neural networks for better performance. Nguyen and Nguyen (2020) Nguyen and Nguyen (2020) developed PhoBERT, a BERT-based monolingual language model for Vietnamese, showing excellence in POS tagging, dependency parsing, named-entity recognition. Trinh et al. (2022) integrated PhoBERT with cross-view training to enhance performance using labeled and unlabeled data.

### 2.2 Treebank Conversion

In the early stages of treebank conversion research, manual and rule-based methods were predominant, especially for widely used languages. These approaches relied on linguists and researchers meticulously developing sets of rules to convert syntactic structures and annotations from one standard to another (Frank, 2001; Avgustinova and Zhang, 2010; Candito et al., 2010). Li et al. (2013) used an annotation adaptation algorithm to convert component dependency tree banks from one annotation guideline to another. Nguyen et al. (2014b) designed a heuristic rule to convert Vietnamese component treebank into dependency treebank. Tran and Miyao (2022) Relying on hand-coded rules to assign categories to constituents convert Universal Dependencies (UD) treebanks to Combinatory Categorial Grammar (CCG) treebanks. However, these rule-based methods faced limitations when applied to languages with fewer resources, like Vietnamese, where treebanks are smaller and annotated data is limited, making it challenging to develop comprehensive rule sets.

In recent years, with the advancement of machine learning technologies, learning-based methods have emerged as a new trend in the field of treebank conversion. These methods, by learning the characteristics of different datasets, are more effective in handling the complexities of languages and data (Truong et al., 2022; Wang et al., 2022). The core of treebank conversion is to find the mapping relationship between the source treebank and the target treebank. Therefore, how to effectively utilize the syntactic information of the source treebank is the key to the problem. Jiang et al. (2018) constructed a supervised transformer by manually annotating a certain scale of dual-tree alignment data and using pattern embedding and TreeLSTM to encode the source-side trees. Wan et al. (2022) adopted a graph-to-graph annotation conversion method, utilizing pseudo data and inherit parameters to guide the graph transformation, achieving an automatic conversion from one annotation standard to another.

## 3 Our Proposed Approach

Treebank conversion aims to convert the source-side tree into the target-side tree. Given the word sequence $S = w_1 w_2 \dots w_n$ of the sentence, the part-of-speech sequence $T = t_1 t_2 \dots t_n$ and the source-side tree $d^{src}$, the task of treebank conversion is to output a syntactic tree $d^{tgt}$ that conforms to the target-side norms. Therefore, effectively utilizing the information from the source-side treebank to guide the treebank conversion is key to the research. Naturally, we consider encoding the source-side treebank and using it as an additional input to the MLP to enhance the model's conversion capability. Specifically, we employ two strategies, Pattern Embedding(PE) and Tree BiLSTM, to encode the source-side tree from both local and global perspectives. Next, we will introduce the main components of the model, including the input layer, encoder layer, and decoder layer.

### 3.1 Embedding

**Input Layer.** We adopt the BiAffine Parser proposed by Dozat and Manning (2017) as our foundational model. As a graph-based dependency parsing model, this parser has achieved outstanding performance across multiple languages. The input layer primarily transforms words in a sentence into dense vectors. Compared to the original parser, we discovered that replacing label representations with character rep-

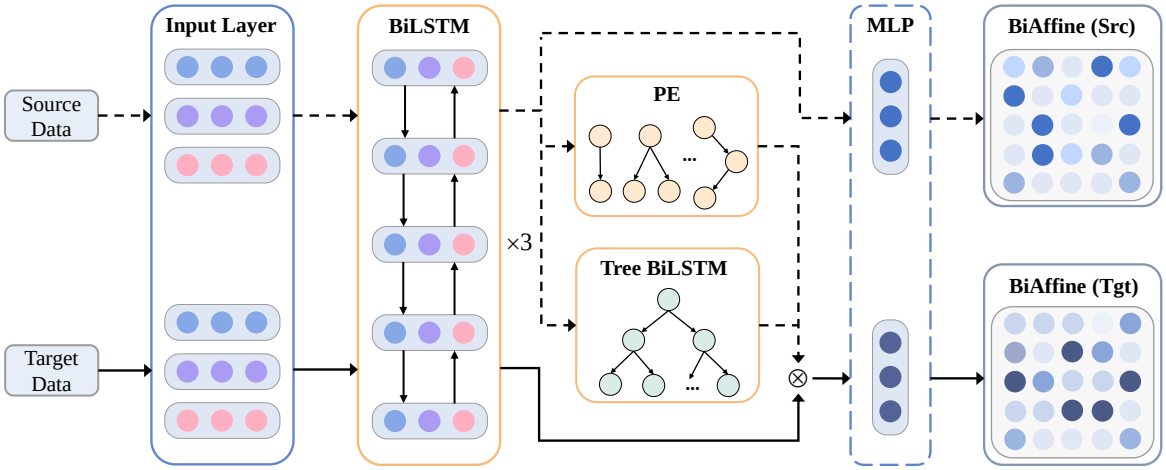



Figure 2: Framework of our proposed model.

resentations could further enhance the model performance. Moreover, to strengthen the representation ability of the model for Vietnamese, we introduced the pre-trained language model XLM-RoBERTa as an additional representation and fine-tuned it using a large volume of unlabeled data to improve its understanding and representation of Vietnamese. Consequently, the final input vector $\mathbf{x}_i$ for the input layer is constructed by concatenating $\mathbf{rep}_i^{\texttt{char}}$, $\mathbf{rep}_i^{\texttt{XLM-R}}$, and $\mathbf{emb}_i^{\texttt{word}}$.

$$\mathbf{x}_i = \mathbf{rep}_i^{\texttt{char}} \oplus \mathbf{rep}_i^{\texttt{XLM-R}} \oplus \mathbf{emb}_i^{\texttt{word}} \tag{1}$$

### 3.2 Encoder

**BiLSTM.** The encoder layer uses a three-layer bidirectional LSTM (BiLSTM) network to obtain a context-sensitive representation of each word. Multi-layer BiLSTM is capable of capturing more complex contextual information and deeper linguistic features. BiLSTM provides a comprehensive representation of each word by combining information from both directions of the sentence.

$$\mathbf{h}_0\mathbf{h}_1 \dots \mathbf{h}_n = \mathbf{BiLSTM}(\mathbf{x}_0\mathbf{x}_1 \dots \mathbf{x}_n, \theta_{\text{BiLSTM}}) \tag{2}$$

**Pattern Embedding.** Li et al. (2012) first proposed a pattern-based treebank conversion method. This method generates pseudo-double tree alignment data by using the syntax model trained by the source-side treebank to parse the target-side treebank. Then, they extracted pattern-based features from the source trees of this pseudo-double-tree alignment data and used these features as additional guidance information to train a more accurate target-side syntax model. However, this definition of pattern features only focuses on syntactic structure information and ignores the dependency relationship of the source tree. Based on this, Jiang et al. (2018) further refined and expanded the pattern features and introduced more label information. Therefore, in order to make full use of the information of the source syntax tree, we adopted 9 detailed pattern classifications based on the work of Jiang et al. (2018), as shown in Table 1.

Taking the calculation of arc $i \leftarrow j$ as an example, first determine the corresponding pattern type $P_{i \leftarrow j}$ based on the positional relationship between word $w_i$ and word $w_j$ in the source tree and refer to Table 1. Then obtain the dependency label corresponding to word $w_i$ and word $w_j$ and put It is represented as $l_i$ and $l_j$. In addition, the closest relationship between word $w_i$ and word $w_j$ is also considered the dependency label $l_a$ of the common ancestor word $a$. Finally, these four embedding representations are merged to form the pattern embedding representation $\mathbf{r}_{i \leftarrow j}^{pat}$.

$$\mathbf{r}_{i \leftarrow j}^{pat} = \mathbf{e}^{P_{i \leftarrow j}} \oplus \mathbf{e}^{l_i} \oplus \mathbf{e}^{l_j} \oplus \mathbf{e}^{l_a} \tag{3}$$

| Target-side arc | Source-side arc | Pattern |
|---|---|---|
| $i \leftarrow j$ | $i \leftarrow j$ | consistent |
| | $i \leftarrow k \leftarrow j$ | grand |
| | $i \leftarrow k \rightarrow j$ | sibling |
| | $i \rightarrow j$ | reverse |
| | $i \rightarrow k \rightarrow j$ | reverse grand |
| | $i \leftarrow k \rightarrow m \rightarrow j, i \leftarrow k \leftarrow m \leftarrow j, \ldots$ | Distance of 3 |
| | $\ldots$ | Distance of 4-5, 6 and $\geq 7$ |

Table 1: Path-Pattern Correspondence Table.

Finally, the vector $\mathbf{r}_{i\leftarrow j}^{pat}$ is concatenated with $h_i$ and $h_j$ respectively, serving as the inputs for $\text{MLP}^{\text{D}}$ and $\text{MLP}^{\text{H}}$.

**Tree BiLSTM.** To better leverage the information from the source-side tree, we naturally thought of using Tree BiLSTM to encode the source-side tree. Tree BiLSTM is an extension of BiLSTM that is more adept at handling tree-structured data. Tree BiLSTM aggregates and transfers information according to the tree structure of the input data, capturing hierarchical information by recursively processing each node in the tree. The state of each node not only depends on its child nodes but may also rely on the states of its parent or sibling nodes, allowing the model to capture long-distance dependencies and complex relationships within the tree structure.

To ensure that the Tree BiLSTM can fully capture the contextual information of the sentence, we directly utilize the output of the top-level BiLSTM as its input. Moreover, to introduce the dependency relations of the source-side tree, we also map the dependency labels of the source-side tree to embedding vectors as an additional input for the Tree BiLSTM. Taking the word $w_k$ as an example, $\mathbf{r_k}$ is its top-level BiLSTM output vector, $l_k$ is the dependency relation label of word $w_k$ in the source tree, and $e^{l_k}$ is the embedding vector mapped from the label. Therefore, the input for the word $w_k$ in Tree BiLSTM is:

$$x_k = \mathbf{r_k} \oplus \mathbf{e}^{l_k} \tag{4}$$

In Tree BiLSTM, the hidden state vector of each node is calculated by combining the input vector of the node with the hidden state vectors of all its child nodes. Let $x_a$ be the input vector for node $a$, and $C_{(a)}$ represent the set of all its child nodes. The hidden state vector $h_k$ and cell state vector $c_k$ correspond to the hidden and cell states of its children, respectively. Thus, the Tree LSTM computation process for node $a$ can be described as follows:

$$\begin{aligned}
\hat{\mathbf{h}}_a &= \sum_{k \in C(a)} \mathbf{h}_k \\
\mathbf{i}_a &= \sigma(\mathbf{W}^{(i)}\mathbf{x}_a + \mathbf{U}^{(i)}\hat{\mathbf{h}}_a + \mathbf{b}^{(i)}) \\
\mathbf{f}_{ak} &= \sigma(\mathbf{W}^{(f)}\mathbf{x}_a + \mathbf{U}^{(f)}\mathbf{h}_k + \mathbf{b}^{(f)}) \\
\mathbf{o}_a &= \sigma(\mathbf{W}^{(o)}\mathbf{x}_a + \mathbf{U}^{(o)}\hat{\mathbf{h}}_a + \mathbf{b}^{(o)}) \\
\mathbf{u}_a &= \tanh(\mathbf{W}^{(u)}\mathbf{x}_a + \mathbf{U}^{(u)}\hat{\mathbf{h}}_a + \mathbf{b}^{(u)}) \\
\mathbf{c}_a &= \mathbf{i}_a \odot \mathbf{u}_a + \sum_{\mathbf{k} \in C(a)} \mathbf{f}_{ak} \odot \mathbf{c}_k \\
\mathbf{h}_a &= \mathbf{o}_a \odot \tanh(\mathbf{c}_a)
\end{aligned} \tag{5}$$

Where $W$ and $U$ are the weight matrices to be learned, $b$ is the bias term, and $\sigma$ is the sigmoid activation function.

After encoding through the Tree BiLSTM, we obtain the top-down word sequence representation $h_k^{\downarrow}$ and the bottom-up word sequence representation $h_k^{\uparrow}$ of the source-side tree. We then concatenate these

CCL 2024

representations, and the output of the Tree BiLSTM is $\mathbf{r}_k^{tree} = h_k^{\downarrow} \oplus h_k^{\uparrow}$. Finally, $\mathbf{r}_k^{tree}$ is used as an additional input for the MLP layer.

### 3.3 Decoder

**MLP layer.** The MLP layer contains two independent multilayer perceptrons (MLP), one for predicting head words in dependencies and the other for predicting dependant words. This layer further merges the output of BiLSTM with the output of PE and Tree BiLSTM and converts it into features for predicting dependencies.

$$\begin{aligned} \mathbf{r}_i^{\mathrm{D}} &= \mathrm{MLP}^{\mathrm{D}}\left(\mathbf{h}_i \oplus \mathbf{r}_{i\leftarrow j}^{pat} \oplus \mathbf{r}_i^{tree}\right) \\ \mathbf{r}_j^{\mathrm{H}} &= \mathrm{MLP}^{\mathrm{H}}\left(\mathbf{h}_j \oplus \mathbf{r}_{i\leftarrow j}^{pat} \oplus \mathbf{r}_j^{tree}\right) \end{aligned} \tag{6}$$

**BiAffine layer.** The BiAffine layer is responsible for calculating the dependency scores between words, where the score represents the dependency arc score from the head word $W_j$ to the dependent word $w_i$.

$$\mathtt{score}(i \leftarrow j) = \begin{bmatrix} \mathbf{r}_i^{\mathrm{D}} \\ 1 \end{bmatrix}^{\mathrm{T}} \mathbf{U}^1 \mathbf{r}_j^{\mathrm{H}} \tag{7}$$

where the dependency arc from $W_j$ to $w_i$ is determined by the weight matrix $\mathbf{U}^1$.

**Parsing loss.** Assuming that the correct core word of word $W_i$ is $W_j$ and the dependency relation label is $l$, the loss function of the model is defined as follows:

$$\begin{aligned} \mathcal{L}^{\mathrm{par}} = &- \log \frac{e^{\mathtt{score}(i\leftarrow j)}}{\sum\limits_{0\le k\le n, k\ne i} e^{\mathtt{score}(i\leftarrow k)}} \\ &- \log \frac{e^{score(i \xleftarrow{l} j)}}{\sum_{l'\in\mathcal{L}} e^{score(i \xleftarrow{l'} j)}} \end{aligned} \tag{8}$$

### 3.4 Multi-task Learning

In order to maximize the use of corpus information on related tasks, we adopt a multi-task learning strategy to achieve the contribution of all corpora to each task by training multiple tasks at the same time. We treat dependency parsing of source-side trees and target-side trees as two independent tasks. These two tasks jointly learn syntactic knowledge by sharing the parameters of the Input Layer and Encoder Layer, while using independent parameter configurations on the MLP layer and BiAffine layer to capture the unique characteristics of their respective specifications. This design not only emphasizes the commonalities between tasks but also retains their unique grammatical features, thereby effectively enhancing the performance of the model on dependency parsing tasks.

## 4 Experiments

### 4.1 Experiment Settings

**Data.** To ensure a fair comparison of the effects of treebank conversion, we selected several published Vietnamese dependency treebanks for our experimental analysis, including UD, VnDT, and VLSP2020. Where UD and VLSP2020 follow the same annotation standards. This chapter primarily discusses the processing and analysis of multiple datasets, including the source treebank VnDT, the target treebank UD, and the bi-tree aligned data $UD^{\mathrm{VnDT}}$. For the bi-tree aligned data, we randomly selected 2900 sentences from VnDT and parsed them using a well-trained target-end dependency analysis model, generating pseudo-aligned tree data with noise. To minimize the impact of annotation errors on model performance, we have manually corrected these data. From the bi-tree aligned data, we randomly selected 300 sentences as the dev set and 600 sentences as the test set. The remaining approximately 2000 sentences were used as the training set. The scale of each dataset is detailed in Table 2.

CCL 2024

| | UD | VLSP2020 | VnDT | $UD^{\text{VnDT}}$ |
|---|---|---|---|---|
| train | 1,400 | 8,152 | 8,977 | 1,996 |
| dev | 800 | - | 200 | 300 |
| test | 800 | 1,123 | 1020 | 600 |

Table 2: Vietnamese dependency treebanks

**Evaluation.** We employed the standard Unlabeled Attachment Score (UAS), which considers only the correctness of dependency arcs, and the Labeled Attachment Score (LAS), which evaluates the accuracy of both dependency arcs and dependency relations, to assess the performance of dependency parsing models.

**Hyper-parameters.** In our model, we use 100-dimensional XLM-RoBERTa pre-trained word vectors and 100-dimensional char embedding vectors. We set the number of layers of the BiLSTM encoding layer to 3, the output dimension to 300, and the dimensions of the MLP layer to 500 and 100. The dimensionality of the pattern embedding vectors is 50, the dimensionality of the dependency relation labels is 50, and the output dimensionality of the Tree BiLSTM is 200.

In the training phase, our batch size is 40 sentences. After training all the training corpus, it is regarded as one iteration. After each iteration, the performance of the model on the development set is evaluated. When the performance of the model does not improve after 50 iterations, training is stop. For the multi-task learning model,we proceed by first training a batch of UD data followed by a batch of VnDT data, considering the completion of one training cycle on VnDT data as one iteration.

### 4.2 Treebank Conversion Results

Table 3 presents the results of the treebank conversion. Given that UD and VnDT adopt completely different annotation standards, we compared the consistency of dependency arcs and dependency labels within $UD^{\text{VnDT}}$, where the consistency of dependency arcs was 52.4%, and the consistency of dependency labels was only 21.3%. Among the two methods we compared, the PE method employs local encoding to encode the source-side trees, while Tree BiLSTM uses global encoding. The experimental results indicate that the encoding approach of Tree BiLSTM achieved the best performance. It shows that the global encoding of the source-side trees can achieve better conversion performance when the data consistency is low. On the contrary, the PE method relies more on data consistency and has certain limitations.

| | $UD^{\text{VnDT}}$ | | | |
|---|---|---|---|---|
| | LAS | UAS | LAS | UAS |
| | with Multi-task | | w/o Multi-task | |
| Tree BiLSTM | **65.88** | **81.06** | 64.80 | 80.09 |
| PE | 62.73 | 77.85 | 62.02 | 76.86 |
| Tree BiLSTM+PE | 63.89 | 78.53 | 63.33 | 79.69 |

Table 3: Conversion accuracy on $UD^{\text{VnDT}}$.

At the same time, it naturally occurred to us to combine the Tree BiLSTM method with the PE methods, with the last row of the table presenting the experimental results. Contrary to our intuition, the experiment combining Tree BiLSTM and PE resulted in a decrease in performance, which we believe is due to information redundancy. As shown in Figure 3, we selected sentences from the test set for analysis. When encoding was done solely with Tree BiLSTM, the parser could produce results perfectly accurately. However, when employing both Tree BiLSTM and PE for encoding, the discrepancy between the UD and VnDT annotation rules was further amplified, leading to erroneous outputs from the model. Although the two methods attempt to encode the source-side syntactic tree from two different perspectives, local and global, the encoding through Tree BiLSTM already contains sufficient syntactic information, resulting in

CCL 2024

redundancy rather than complementarity of the final syntactic information.

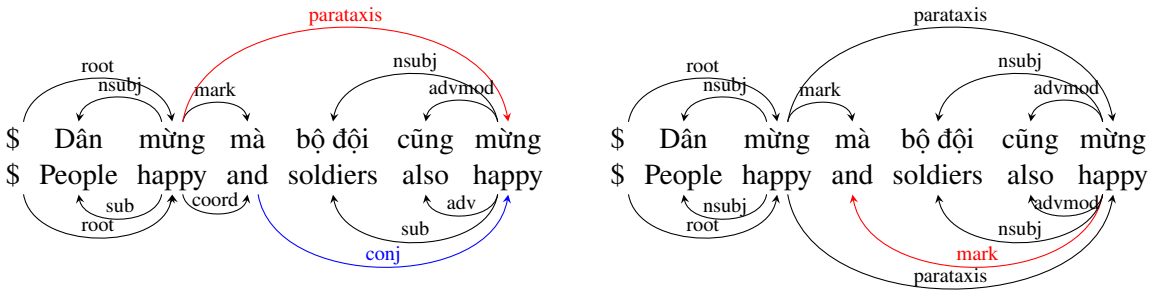



Figure 3: On the left, the upper half shows the target-side UD tree, and the under half shows the source-side VnDT tree. On the right, the upper half shows the output after Tree BiLSTM encoding, and the under half shows the output after combining Tree BiLSTM and PE.

### 4.3 Analysis

**Ablation study.** As the syntactic information fused in the model comprises a combination of multiple syntactic features, we conducted ablation experiments on these features. The results in table 4 illustrate the effects of removing fused source-side syntactic features from two different methods. For the PE method, the removal of any feature resulted in a slight decrease in model performance, indicating that incorporating source-side syntactic information positively impacts the model. The most significant impact on the UAS metric was the removal of $l_i$, which represents the dependency relation between word $w_i$ and its head word, with a decrease of 1.52%. This suggests that source-side dependency relations indeed assist in treebank conversion. For the Tree BiLSTM methods, introducing dependency relation labels from the source-side trees slightly improved the model's conversion effectiveness. After integrating the source-side trees, LAS improved by 1.69% and UAS improved by 1.49%.

| | $UD^{\text{VnDT}}$ | | | |
|---|---|---|---|---|
| | LAS | UAS | LAS | UAS |
| | with Multi-task | | w/o Multi-task | |
| PE | **62.73** | **77.85** | 62.02 | 76.86 |
| w/o $l_i$ | 61.38 | 76.33 | 61.03 | 76.11 |
| w/o $l_j$ | 61.47 | 76.45 | 61.14 | 76.20 |
| w/o $l_a$ | 60.89 | 76.82 | 60.52 | 76.24 |
| TreeLSTM | **65.88** | **81.06** | 64.80 | 80.90 |
| w/o labels | 64.81 | 80.16 | 64.19 | 79.57 |

Table 4: Feature ablation.

**Error analysis.** Inspired by the ablation experiments, we delved deeper into the effects of label types and both absolute and relative distances on conversion performance of the model. As shown in Figure 4, we analyzed the top ten most frequent label types in the dataset and their respective accuracy rates. The results indicated that the higher the frequency of occurrence for a label, the higher the accuracy of conversion, aligning with our understanding. However, some labels deviated from this pattern. We believe this discrepancy results from significant differences in the annotation standards of the dataset. For instance, the label "nsubj" under the UD dataset's annotation standards can be further subdivided into "nsubj", "nsubj:nn", "nsubj:pass","nsubj:xsubj" etc., which could significantly impact conversion accuracy of the model. Furthermore, we also evaluated the consistency between the UD and VnDT datasets, revealing label consistency at only 52.4% and arc consistency at 21.3%. This further validates the significant impact of annotation standard differences on the model's conversion performance.

CCL 2024

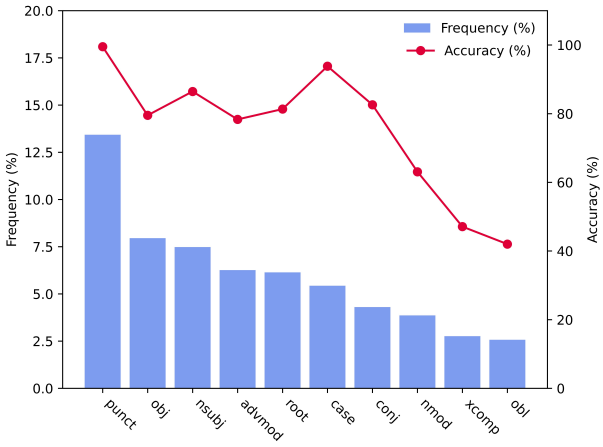



Figure 4: Statistics on the accuracy of the distances.

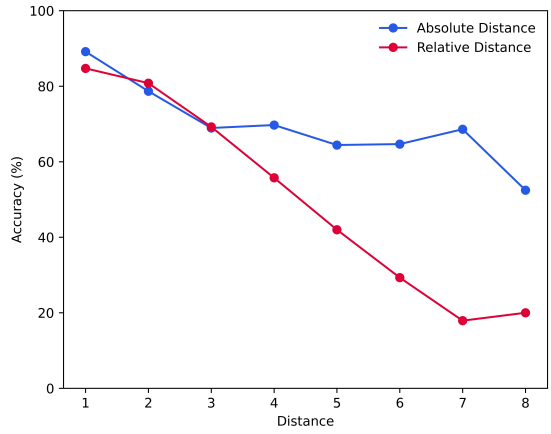



Figure 5: Statistics on the frequency and accuracy of the dependency labels.

In addition, we also analyzed the impact of both relative and absolute distances between words in a sentence on accuracy, as illustrated in Figure 4. Regardless of whether it was relative or absolute distance, accuracy tended to decrease as the distance increased. However, interestingly, when the relative distance was 8, there was a slight increase in accuracy. Upon analysis, we believe this anomaly is due to the fewer instances of dependency arcs with a relative distance of 8 in the dataset. Hence, we counted the number of dependency arcs with relative distances ranging from 1 to 8 in the dual-tree aligned data $UD^{\text{VnDT}}$. We found that there were 232 dependency arcs at a distance of 6, and only 10 at a distance of 8. Therefore, the reduction in the number of instances to some extent could cause a slight increase in accuracy.

A similar situation exists for absolute distance. As shown in Figure 5, when the absolute distance is 7, the accuracy rate is improved by 3.83% compared with when the absolute distance is 6. We found that at an absolute distance of 6, a total of 24 different labels appeared; While at an absolute distance of 7, this number was reduced to 18. Further analysis found that “punct" is the most frequently occurring tag among all tag types. Specifically when the absolute distance is 6, the proportion of “punct" label is 31%, and when the absolute distance is 7, this proportion rises to 40%. From the analysis in Figure 4, we can see that the “punct" is the label with the highest frequency of occurrence and the highest accuracy. The high proportion of “punct" labels is directly related to the improvement of its accuracy. At the same time, the reduction of label types also helps to improve accuracy to a certain extent.

### 4.4 Utilization of Conversion Data

Table 5 presents the final experimental results utilizing the converted data. Choudhary and O’riordan (2023) employ BERT fine-tuned in 12 languages as an encoder on an End-to-end Seq2seq Dependency Parser. Le-Hong and Cambria (2024) integrate dependency graph embeddings into the model using GRU. Compared with these works, our model achieves excellent results even without additional heterogeneous data, further verifying the effectiveness of our model.

| Model | Training Data | LAS | UAS |
|---|---|---|---|
| Choudhary and O’riordan (2023) | UD | - | 68.37 |
| Le-Hong and Cambria (2024) | UD | 47.24 | 52.06 |
| Our Model | UD | **65.10** | **78.38** |
| Multi-task | UD&VnDT | 65.25 | 78.65 |
| Our Model | UD&Converted VnDT | 72.22 | 82.23 |

Table 5: Performance of different dependency parsing models on the test set.

We test the base model, a multi-task learning model, and a conversion model employing the Tree BiLSTM method on the UD dataset. The results indicate that compared to the base model, the multi-task learning model did not demonstrate significant performance improvements. This may be due to the low consistency between the source treebank and the target treebank, and the two share fewer features, resulting in insufficient improvement in model performance. Furthermore, we use the trained conversion model to convert the VnDT training set, and then merge the converted VnDT treebank and UD treebank as a large training set to train the model. It can be seen that as the data size increases, the performance of the model has been greatly improved. LAS and UAS have improved by 7.12% and 3.85% respectively, which proves that the treebank conversion method can effectively utilize the syntactic information of the source treebank. In summary, the treebank conversion method that introduces source-side syntactic information can effectively improve the performance of the target-side syntactic analysis model; and when the consistency of the data is low, compared with the multi-task learning method, the treebank conversion method can be more direct and efficient Utilize the source treebank.

## 5 Conclusion

In this work, we presented a multi-feature enhanced multi-task learning framework for Vietnamese treebank conversion, addressing the challenges of limited corpora and the scarcity of large-scale annotated data in Vietnamese dependency parsing. Our approach leverages Tree BiLSTM and pattern embedding to capture both global and local syntactic features from the source treebank, and integrates these features into a multi-task learning framework, enhancing the representation capability for Vietnamese. The experiments conducted on benchmark datasets demonstrate the effectiveness of our model, which not only facilitates the conversion of heterogeneous Vietnamese Treebanks into a unified one but also significantly improves the Vietnamese dependency parsing accuracy.

## Acknowledgements

This work was supported by the National Natural Science Foundation of China (U21B2027, 62366027, 62266028, 62306129), Yunnan Provincial Major Science and Technology Special Plan Projects (202103AA080015, 202202AD080003, 202203AA080004), Yunnan Fundamental Research Projects (202401BC070021, 202301AS070047, 202401CF070121), Kunming University of Science and Technology's "Double First-rate" Construction Joint Project (202201BE070001-021, 202301BE070001-027), Yunnan High and New Technology Industry Project (201606). The authors would like to thank the anonymous reviewers for their valuable comments and suggestions for improving this paper.

## References

Sharefah Al-Ghamdi, Hend Al-Khalifa, and Abdulmalik Al-Salman. 2023. Fine-tuning bert-based pre-trained models for arabic dependency parsing. *Applied Sciences*, 13(7).

Afra Amini, Tianyu Liu, and Ryan Cotterell. 2023. Hexatagging: Projective dependency parsing as tagging. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1453–1464, Toronto, Canada, July. Association for Computational Linguistics.

órunn Arnardóttir, Hinrik Hafsteinsson, Einar Freyr Sigursson, Kristín Bjarnadóttir, Anton Karl Ingason, Hildur Jónsdóttir, and Steinór Steingrímsson. 2020. A Universal Dependencies conversion pipeline for a Penn-format constituency treebank. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 16–25, Barcelona, Spain (Online), December. Association for Computational Linguistics.

Tania Avgustinova and Yi Zhang. 2010. Conversion of a russian dependency treebank into hpsg derivations. In *Ninth International Workshop on Treebanks and Linguistic Theories*, page 7.

Emanuel Borges Völker, Maximilian Wendt, Felix Hennig, and Arne Köhn. 2019. HDT-UD: A very large Universal Dependencies treebank for German. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 46–57, Paris, France, August. Association for Computational Linguistics.

Marie Candito, Benoît Crabbé, and Pascal Denis. 2010. Statistical french dependency parsing: treebank conversion and first results. In *Seventh International Conference on Language Resources and Evaluation-LREC 2010*, pages 1840–1847. European Language Resources Association (ELRA).

Chinmay Choudhary and Colm O'riordan. 2023. Multilingual end-to-end dependency parsing with linguistic typology knowledge. In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 12–21.

Duc Do, Dien Dinh, An-Vinh Luong, and Thao Do. 2022. Adapting cross-lingual model to improve vietnamese dependency parsing. In *Artificial Intelligence in Data and Big Data Processing*, pages 97–108, Cham. Springer International Publishing.

Timothy Dozat and Christopher D. Manning. 2017. Deep biaffine attention for neural dependency parsing. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Anette Frank. 2001. Treebank conversion. converting the negra treebank to an ltag grammar. In *Proceedings of the Workshop on Multi-layer Corpus-based Analysis*, pages 29–43.

Xinzhou Jiang, Zhenghua Li, Bo Zhang, Min Zhang, Sheng Li, and Luo Si. 2018. Supervised treebank conversion: Data and approaches. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2706–2716.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing universal dependencies universally.

Phuong Le-Hong and Erik Cambria. 2024. Integrating graph embedding and neural models for improving transition-based dependency parsing. *Neural Computing and Applications*, 36(6):2999–3016.

Zhenghua Li, Ting Liu, and Wanxiang Che. 2012. Exploiting multiple treebanks for parsing with quasi-synchronous grammars. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–684.

Xiang Li, Wenbin Jiang, Yajuan Lü, and Qun Liu. 2013. Iterative transformation of annotation guidelines for constituency parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–596.

Zuchao Li, Jiaxun Cai, Shexia He, and Hai Zhao. 2018. Seq2seq dependency parsing. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3203–3214, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Zuchao Li, Hai Zhao, and Kevin Parnow. 2020. Global greedy dependency parsing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8319–8326, Apr.

Zuchao Li, Kevin Parnow, and Hai Zhao. 2022. Incorporating rich syntax information in grammatical error correction. *Information Processing Management*, 59(3):102891.

Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1037–1042, Online, November. Association for Computational Linguistics.

Dat Quoc Nguyen, Dai Quoc Nguyen, Son Bao Pham, Phuong-Thai Nguyen, and Minh Le Nguyen. 2014a. From treebank conversion to automatic dependency parsing for vietnamese. In *Natural Language Processing and Information Systems*, pages 196–207, Cham. Springer International Publishing.

Dat Quoc Nguyen, Dai Quoc Nguyen, Son Bao Pham, Phuong-Thai Nguyen, and Minh Le Nguyen. 2014b. From treebank conversion to automatic dependency parsing for vietnamese. In *Natural Language Processing and Information Systems: 19th International Conference on Applications of Natural Language to Information Systems, NLDB 2014, Montpellier, France, June 18-20, 2014. Proceedings 19*, pages 196–207. Springer.

Dongqi Pu and Khalil Sima'an. 2022. Passing parser uncertainty to the transformer: Labeled dependency distributions for neural machine translation. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 41–50, Ghent, Belgium, June. European Association for Machine Translation.

Sara Stymne, Miryam de Lhoneux, Aaron Smith, and Joakim Nivre. 2018. Parser training with heterogeneous treebanks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 619–625, Melbourne, Australia, July. Association for Computational Linguistics.

Nguyen Duc Thien, Nguyen Thi Thu Trang, and Truong Dang Quang. 2020. Applying graph neural networks for Vietnamese dependency parsing. In *Proceedings of the 7th International Workshop on Vietnamese Language and Speech Processing*, pages 54–59, Hanoi, Vietnam, December. Association for Computational Lingustics.

CCL 2024

Tu-Anh Tran and Yusuke Miyao. 2022. Development of a multilingual ccg treebank via universal dependencies conversion. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5220–5233.

Duc Huu Trinh, Trinh Le-Phuong Ngo, Long Hong Buu Nguyen, and Dien Dinh. 2022. Applying cross-view training for dependency parsing in vietnamese. *ICIC express letters. Part B, Applications: an international journal of research and surveys*, 13(3):249–259.

Chau M. Truong, Tai V. Pham, Minh N. Phan, Nhan D. T. Le, Thinh V. Nguyen, and Quy T. Nguyen. 2022. Converting a constituency treebank to dependency treebank for vietnamese. In *2022 RIVF International Conference on Computing and Communication Technologies (RIVF)*, pages 256–261.

Yuxuan Wang, Zhilin Lei, Yuqiu Ji, and Wanxiang Che. 2022. Simple and effective graph-to-graph annotation conversion. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5450–5460, Gyeongju, Republic of Korea, October. International Committee on Computational Linguistics.

Songlin Yang and Kewei Tu. 2022. Headed-span-based projective dependency parsing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2188–2200, Dublin, Ireland, May. Association for Computational Linguistics.

Yu Zhang, Qingrong Xia, Shilin Zhou, Yong Jiang, Guohong Fu, and Min Zhang. 2022. Semantic role labeling as dependency parsing: Exploring latent tree structures inside arguments. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4212–4227, Gyeongju, Republic of Korea, October. International Committee on Computational Linguistics.

CCL 2024

# SimCLNMT: A Simple Contrastive Learning Method for Enhancing Neural Machine Translation Quality

**Menglong Xu** **Yanliang Zhang***
School of Physics & Electronic Information Engineering, Henan Polytechnic University
Henan Jiaozuo 454000, China
xml@home.hpu.edu.cn ylzhang@hpu.edu.cn

## Abstract

Neural Machine Translation (NMT) models are typically trained using Maximum Likelihood Estimation (MLE). However, this approach has a limitation: while it might select the best word for the immediate context, it does not generally optimize for the entire sentence. To mitigate this issue, we propose a simple yet effective training method called **SimCLNMT**. This method is designed to select words that fit well in the immediate context and also enhance the overall translation quality over time. During training, **SimCLNMT** scores multiple system-generated (candidate) translations using the logarithm of conditional probabilities. It then employs a ranking loss function to learn and adjust these probabilities to align with the corresponding quality scores. Our experimental results demonstrate that **SimCLNMT** consistently outperforms traditional MLE training on both the NIST English-Chinese and WMT'14 English-German datasets. Further analysis also indicates that the translations generated by our model are more closely aligned with the corresponding quality scores. We release our code at https://github.com/chaos130/fairseq_SimCLNMT.

## 1 Introduction

Neural machine translation (NMT) (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Bahdanau et al., 2015) is a pivotal task within the field of natural language processing. It focuses on translating text from one language (source) to another (target) using end-to-end neural network models. This task has garnered increasing attention in both academia and industry (Shen et al., 2016; Vaswani et al., 2017; Kreutzer et al., 2018; Yang et al., 2018). Notably, the emergence of the Transformer architecture (Vaswani et al., 2017) marks a paradigm shift in NMT. Characterized by its innovative self-attention mechanism, the Transformer enables the model to variably prioritize the significance of words within the input sentence. This results in more effectively capturing the context compared to previous models, such as the Recurrent Neural Network (RNN) (Bahdanau et al., 2015) and Convolutional Neural Network (CNN) (Gehring et al., 2017). Despite the advancements of the Transformer, it still relies on the Maximum Likelihood Estimation (MLE) for word prediction at each time step. While this practice optimizes for immediate word selection probability, it may not consistently enhance the overall translation quality. This discrepancy between probabilistic optimality and actual translation quality highlights a critical limitation in current NMT approaches. To assess the practical impact of this limitation, we conducted a preliminary study using a pre-trained NMT model (MLE baseline). Our process involves generating two candidate translations for each input and analyzing whether higher probability correlates with higher BLEU scores (Papineni et al., 2002). As depicted in Table 1, the results indicate that the accuracy is far from ideal, suggesting that there is room for improvement in how NMT models select words for translation.

As suggested by Kreutzer et al. (2018), one primary strategy to mitigate the mentioned issues involves adopting Reinforcement Learning (RL). RL training allows models to be fine-tuned with reward mechanisms that assess the overall quality of predictions. This technique focuses on the result of the translation

*Corresponding author
©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1047-1058, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers

process rather than just the immediate predictions of the next word. However, it introduces common challenges to deep RL, including noisy gradient estimation (Greensmith et al., 2004), which can result in unstable training and high sensitivity of hyperparameter settings. Moura Ramos et al. (2023) further explores this concept by incorporating evaluation metrics learned from human quality annotations as reward models within the NMT framework. This approach enhances translation quality through data filtering, RL training, and re-ranking techniques. As an alternative strategy, Minimum Risk Training (MRT) has been adopted for language generation tasks to overcome the MLE training challenge (Shen et al., 2016). Despite its potential, the accuracy of the MRT is limited by the number of output samples. Yang et al. (2018) introduces a novel method leveraging Generative Adversarial Network (GAN) for NMT, aiming to train the model to generate sentences indistinguishable from human translations. Additionally, Liu et al. (2022) and Yuan et al. (2023) have demonstrated that contrastive learning techniques can effectively improve the quality of model output.

| Dataset(dev) | Bleu | Acc(%) |
|---|---|---|
| NIST en-ch | 45.61 | 58.54 |
| WMT14 en-de | 25.93 | 54.10 |

Table 1: Accuracy of pre-trained NMT models w.r.t ranking the quality of candidate translations on NIST en-ch dataset (dev) and WMT14 en-de (dev). **Acc.** stands for the frequency of the model that assigns higher probabilities to better candidate translations. The candidate translations are generated by a pre-trained model. **Bleu** are the evaluation metric scores in the development data set. The pre-trained model only achieves 58.54% and 54.10% precision on both development data sets.

In this paper, we seek to address the mentioned issues by incorporating contrastive learning into the NMT framework. Inspired by Liu et al. (2022) and Yuan et al. (2023), we introduce **SimCLNMT**, a simple yet effective approach that aligns the probabilities of various candidate translations with their corresponding quality scores via a ranking loss mechanism. Our methodology is divided into two phases: pre-training and fine-tuning. During the pre-training phase, we train a NMT model using a comprehensive data set, which serves as a strong MLE baseline for the fine-tuning stage. In the fine-tuning phase, we generate several candidate translations from the MLE baseline model. Each candidate is then evaluated using an automated metric to determine its quality score, yielding pairs of candidate translations and quality scores for every source sentence. **SimCLNMT** computes the log probability for each candidate and uses a ranking loss mechanism to align these probabilities with their corresponding quality scores, ensuring that the model predictions are consistent with the evaluations provided by the metric. In summary, we mainly make the following contributions:

- To the best of our knowledge, this work is among the first efforts to leverage contrastive learning as a strategy to mitigate the limitations inherently caused by the MLE training in the context of Neural Machine Translation.
- Our model shows a significant performance enhancement when either BLEU or COMET is used as the designated evaluation metric for ranking candidate translations. This suggests that our approach can optimize any desired target metric and serve as an alternative to RL and MRT.
- We conduct extensive experiments on NIST English-Chinese and WMT'14 English-German datasets, and the results show that our approach can achieve consistent improvements.

## 2 Related Work

The training of NMT models typically relies on MLE. While MLE excels at selecting contextually appropriate words, it often falls short of optimizing overall sentence performance, which can result in suboptimal translation outputs. Several previous works have aimed to mitigate this issue. Kreutzer et al. (2018) explores the use of Reinforcement Learning (RL) to improve machine translation with human feedback. This approach fine-tunes NMT models using RL techniques and rewards that assess the overall

quality of translations, rather than focusing solely on predicting the next word. However, it also faces the inherent challenges of deep RL, including the problem of noisy gradient estimation. Moura Ramos et al. (2023) extends the exploration of using RL for NMT by integrating evaluation metrics derived from human feedback as reward models within the NMT framework. This methodology encompasses multiple strategies that include data filtering to refine the training corpus, RL training to adjust the NMT model based on rewards reflecting translation quality, and re-ranking techniques at inference time. This comprehensive approach seeks to enhance translation quality by ensuring that the outputs of the NMT model are more closely aligned with human judgments. Shen et al. (2016) presents a significant shift in NMT training methodologies by incorporating evaluation metrics as loss functions. This allows the model to learn patterns that better align with the quality of the final translation output, rather than just the likelihood of word sequences. Yang et al. (2018) introduces a novel method that utilizes GAN for NMT. In this approach, the NMT model acts as the generator, producing human-like translations, while a discriminator is trained to distinguish between machine and human translations. The goal is to iteratively improve the NMT model based on discriminator feedback, training it to generate translations indistinguishable from human ones and directly optimizing the output quality.

In addition to the above methodologies, contrastive learning has gained traction, particularly in abstractive summarization and large language models. Liu and Liu (2021) and Liu et al. (2022) introduce contrastive learning to the domain of abstractive summarization, focusing on learning from higher-quality sentences rather than those with high model probabilities. This approach shifts the focus towards generating summaries that are statistically likely and of higher relevance and quality. Yuan et al. (2023) further refines the integration of human feedback into language model fine-tuning through a novel training paradigm called RRHF. This method evaluates sampled responses from various sources using the logarithm of conditional probabilities and employs a ranking loss mechanism to align these probabilities with human preferences.

## 3 SimCLNMT

This section breaks down the holistic training process into two stages: pretraining (Section 3.1) and finetuning (Section 3.2).

### 3.1 Pretraining

In this phase, our goal is to pre-train a MLE based model $\pi$ that results in good translations. MLE is the standard training method. It aims to maximize the likelihood of the reference translation $y$, i.e.,

$$\theta^* = \underset{\theta}{\operatorname{argmax}} \sum_i \log p_{\pi_\theta} \left( y^{*(i)} \mid x^{(i)}; \theta \right) \tag{1}$$

where $\theta$ denotes the parameters of $\pi$ and $p_{\pi_\theta}$ denotes the probability distribution entailed by these parameters. The summation is over the training set and $\left\{x^{(i)}, y^{*(i)}\right\}$ is the $i$-th training sample.

For a specific sample $\left\{x^{(i)}, y^{*(i)}\right\}$, Eq. 1 is equal to minimizing the sum of negative log-likelihoods of the tokens $\left\{y_1^*, \ldots, y_j^*, \ldots, y_l^*\right\}$ in the reference translation $y^*$ whose length is $l$, which is the cross-entropy loss:

$$L_{\text{xent}} = -\sum_{j=1}^{l} \sum_y p_{\text{true}} \left( y \mid x, y_{<j}^* \right) \log p_{\pi_\theta} \left( y \mid x, y_{<j}^*; \theta \right) \tag{2}$$

where $y_{<j}^*$ denotes the preceding reference sequence. In practice, $p_{\text{true}}$ is a label-smoothing distribution under the standard MLE framework by assigning probability mass $\beta$ to other tokens:

$$p_{\text{true}} \left( y \mid x, y_{<j}^* \right) = \begin{cases} 1-\beta & y = y_j^* \\ \frac{\beta}{N-1} & y \neq y_j^* \end{cases} \tag{3}$$

### 3.2 Finetuning

In this section, we first introduce the methods of candidate generation and evaluation. Subsequently, the section provides a detailed explanation of the ultimate training objectives of the proposed model.

CCL 2024

**Candidate Generation** NMT typically uses beam search (Freitag and Al-Onaizan, 2017) or diverse beam search (Vijayakumar et al., 2016) to generate candidate translations. Both are approximate algorithms to identify the maximum a-posteriori (MAP) output, i.e., the sentence with the largest estimated probability given an input (Edunov et al., 2018a). Beam search generally finds the high probability outputs (Ott et al., 2018).

However, MAP prediction can lead to less rich translations (Ott et al., 2018) since it aims to select the output sequence with the highest probability given the input. As a result, the model may generate accurate translations, but they may lack the nuances and unique expressions present in the source text.

As an alternative, we consider sampling-based approaches for generating candidate translations. This approach allows the model to explore less frequent, potentially more interesting, or contextually appropriate translations that deterministic methods (beam search) might overlook. First, we consider exploring pure sampling, which leads to higher variability in generation but may also result in less coherent translations due to the random selection of words with lower probabilities. Second, we employ Top-k sampling, which limits the sampling pool to the $k$ most probable words at each time step. This reduces the chance of selecting highly improbable words, leading to more coherence while maintaining some level of diversity (Graves, 2013; Ott et al., 2018; Fan et al., 2018).

**Candidate Evaluation** Evaluation metrics are indispensable tools in the field of machine translation. They serve two critical functions: measuring the quality of translations produced by a model and guiding the iterative improvement of the NMT system through training. Previous work predominantly use BLEU (Papineni et al., 2002) to guide the training of the NMT system (Shen et al., 2016; Yang et al., 2018). However, BLEU relies on word overlap and n-gram matching, making them ineffective for translations that have the same meaning but are substantially different from the reference. To address these shortcomings, we use robust neural metrics to evaluate candidate translations, such as reference-based COMET (Rei et al., 2022).

**Training Objective** After the two mentioned stages, we engage $n$ candidate-translation and quality-score pairs $(y_i, r_i)$ for a source sentence $x$, in which $r_i$ equals the evaluation metrics score $M(x, y_i)$. In this context, $n$ serves as a hyperparameter. As the Figure 1 shows, to align with the quality scores $\{r_i\}_k$, our pre-trained NMT model $\pi$ generates the model scores $p_i$ for every $y_i$, according to the following formula:

$$p_i = \frac{\sum_t \log P_\pi \left(y_{i,t} \mid x, y_{i,<t}\right)}{||y_i||^\alpha} \tag{4}$$

Here, $p_i$ denotes the conditional logarithmic probability normalized in the length of $y_i$ in the model $\pi$, with $\alpha$ as the length normalization coefficient. Our approach is straightforward: we configure the model $\pi$ to assign higher probabilities to the superior candidate translations and lower probabilities to the inferior ones. To refine model performance towards this goal, we employ a ranking loss:

$$L_{\text{rank}} = \sum_{i=1}^{N} \sum_{j=1}^{N} \theta(r_i < r_j) \cdot \max(0, p_i - p_j) \tag{5}$$

This equation introduces the indicator function $\theta(r_i < r_j)$, which equals 1 when the condition $r_i < r_j$, and 0 otherwise.

Following Edunov et al. (2018b), we combine the ranking loss (Eq. 5) with the cross-entropy (Eq. 2) losses to enhance the generative ability of the pre-trained NMT model. The total loss, as defined by (Eq. 6), is the weighted sum of these two losses.

$$L = L_{\text{rank}} + \lambda L_{xent} \tag{6}$$

## 4 Experiments and Results

We evaluate our **SimCLNMT** on the NIST English-Chinese and the WMT'14 English-German translation tasks. We experiment **SimCLNMT** on transformer-base model.

CCL 2024

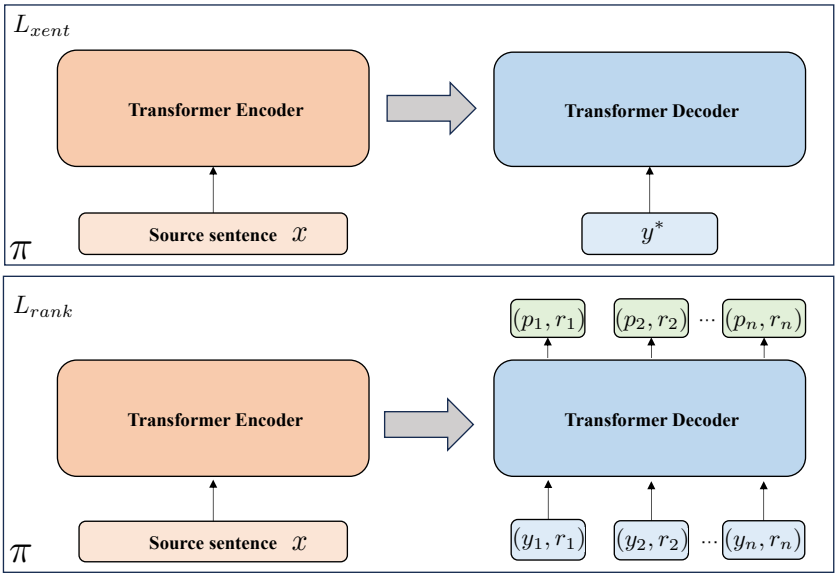



Figure 1: The $\pi$ refers to the model trained in the pre-training phase. $y^*$ denotes the reference translation of source sentence $x$

### 4.1 Data Sets and Preprocessing

**English-Chinese** For English-Chinese translation, our training data comprises 1.6 million sentence pairs. These are randomly extracted from LDC corpora[1]. Both the source and target sentences are encoded with byte-pair encoding (BPE) (Sennrich et al., 2016). The vocabulary for the source and target language consists of approximately 32000 tokens respectively[2]. We choose the NIST02 dataset as our development set, which includes 878 sentences. For testing, we utilize the NIST03, NIST04, NIST05, and NIST06 datasets, which contain 919, 1788, 1082, and 1664 sentences, respectively.

**English-German** In our English-German translation experiments, we use the WMT'14 English-German corpus. This publicly available dataset is widely recognized as a benchmark for NMT systems and contains 4.5 million sentence pairs[3]. Sentences are encoded with BPE, which employs a shared source-target vocabulary of about 37000 tokens. We report results on newstest2014. The newstest2013 is used as validation. These datasets contain 3003 and 3000 sentences, respectively.

### 4.2 Model Parameters and Evaluation

For the Transformer model, following the base model in Vaswani et al. (2017), we set the dimension of the word embedding as 512, the dropout rate as 0.1, and the number of attention heads to 8. Both the encoder and decoder have a stack of 6 layers. During testing, We use beam search with a beam size of 4 and a length penalty of 0.6.

We reimplement the Transformer model in Pytorch using the fairseq (Ott et al., 2019) toolkit[4] and train it synchronously across four GTX 3090 GPUs in a multi-GPU setup on a single machine. To assess the performance of each NMT system, we utilize well-established evaluation metrics, including case-sensitive detokenized BLEU with SacreBLEU [5] (Post, 2018) and ChrF (Popović, 2015).

[1]LDC2002L27, LDC2002T01, LDC2002E18, LDC2003E07, LDC2004T08, LDC2004E12, LDC2005T10

[2]When doing BPE for Chinese, we need to do word segmentation first using jieba and the following steps are the same with BPE for English.

[3]http://nlp.stanford.edu/projects/nmt

[4]Code available at https://github.com/pytorch/fairseq

[5]https://github.com/mjpost/sacrebleu

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1047-1058, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

| Model | Metric | English-Chinese | | | | | Enlish-German |
|---|---|---|---|---|---|---|---|
| | | NIST03 | NIST04 | NIST05 | NIST06 | average | newstest2014 |
| MLE | BLEU | 39.05 | 37.53 | 36.18 | 37.31 | 37.52 | 26.21 |
| | ChrF | 33.81 | 32.39 | 31.50 | 32.27 | 32.49 | 57.19 |
| **SimCLNMT** | BLEU | 39.79 | 38.56 | 36.86 | 38.57 | **38.45** | **26.84** |
| **(diverse beam search)** | ChrF | 34.10 | 32.96 | 31.75 | 33.03 | **32.96** | **57.85** |
| SimCLNMT | BLEU | 39.66 | 38.11 | 36.97 | 38.59 | 38.33 | 26.64 |
| (pure sampling) | ChrF | 33.95 | 32.67 | 31.77 | 32.96 | 32.84 | 57.60 |
| SimCLNMT | BLEU | 39.81 | 38.16 | 37.04 | 38.39 | 38.35 | 26.74 |
| (top-10 Sampling) | ChrF | 34.08 | 32.63 | 31.87 | 32.83 | 32.85 | 57.61 |

Table 2: BLEU and ChrF score on English-Chinese and English-German translation tasks. **diverse beam search, pure sampling, top-10 sampling** represents different methods for generating candidate translations in Section 3.2.

### 4.3 Training Details

During the fine-tuning phase, We randomly select 1,000 examples from the NIST training set and 2000 examples from the WMT14 training set. To ensure the reliability and broad applicability of our findings, we use three different random seeds for the sample selection process and report the average results from the test set. To generate candidate translations for each source sentence, we employ various methods, including diverse beam search, pure sampling, and Top-10 sampling. Our experiments indicate that the hyper-parameter $n$ had a negligible impact on the results. Consequently, based on the performance of the development set, we set the value of $n$ to 32. We then fine-tune the pre-trained NMT model using the Adam optimizer, adjusting the learning rate as necessary. We train the model for fifteen epochs, selecting the best-performing model on the development set for further evaluation. For more details, please refer to Appendix A.

### 4.4 Main Results

Table 2 presents the main results of English-Chinese and English-German test sets. Employing diverse beam search, our model (i.e., the line of SimCLNMT (diverse beam search) in Table 2) achieves an average improvement of +0.93 points in BLEU and +0.47 points in ChrF for the English-Chinese test set. For the English-German test set, the gains are +0.63 points in BLEU and +0.66 points in ChrF. Using pure sampling, the model (the line of SimCLNMT (pure sampling)) shows improvements up to +0.81 points in BLEU and +0.35 points in ChrF on the English-Chinese test set. For the English-German test set, the gains are +0.43 points in BLEU and +0.41 points in ChrF. With top-10 sampling, the model (the line of SimCLNMT (top-10 sampling)) achieves an average improvement of +0.83 points in BLEU and +0.36 points in ChrF on the English-Chinese test set. For the English-German test set, the gains are +0.53 points in BLEU and +0.42 points in ChrF.

These findings demonstrate that our proposed model consistently outperforms MLE baseline, with optimal results observed when using diverse beam search. Notably, when employing sampling-based methods to generate candidate translations, the performance of our model does not decline significantly compared to diverse beam search. This robustness across different data distributions underscores the utility and effectiveness of our model.

## 5 Analysis

### 5.1 Ablation Study

In our investigation, we perform an ablation study to assess the effectiveness of the **SimCLNMT** approach. We train two distinct models under identical conditions, with the only difference being the loss function used. The first model is trained using the Maximum Likelihood Estimation (MLE) method, with the loss function as defined in Equation 2. In contrast, the second model integrates the **SimCLNMT**

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1047-1058, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

technique, using the loss function introduced in Equation 5. As shown in Table 3, the **SimCLNMT(5)** model exhibits a significant improvement over the **SimCLNMT(2)** model. This comparison underscores the effectiveness of our proposed method. In the subsequent subsection, we uniformly employ diverse beam search to generate candidate translations across all our experiments.

| Model | Metric | English-Chinese | | | | | English-German |
|---|---|---|---|---|---|---|---|
| | | NIST03 | NIST04 | NIST05 | NIST06 | average | Newstest2014 |
| SimCLNMT(2) | BLEU | 39.00 | 37.57 | 36.08 | 37.37 | 37.51 | 26.20 |
| | ChrF | 33.76 | 32.40 | 31.40 | 32.28 | 32.46 | 57.26 |
| **SimCLNMT(5)** | BLEU | 39.87 | 38.51 | 36.80 | 38.71 | **38.47** | **26.84** |
| | ChrF | 34.15 | 32.90 | 31.69 | 33.14 | **32.97** | **57.80** |

Table 3: SimCLNMT(2) is trained using the loss function presented in Equation 2, whereas **SimCLNMT(5)** utilizes the loss function detailed in Equation 5. These results focus on the candidate translations obtained through diverse beam search.

| Model | BLEU | ChrF | COMET |
|---|---|---|---|
| MLE | 37.52 | 32.46 | 84.99 |
| **SimCLNMT(B)** | **38.45** | **32.96** | 85.18 |
| SimCLNMT(C) | 38.46 | 32.95 | **85.50** |

a

Table 4: Results on NIST English-Chinese using different evaluation metrics as M in Section 3.2. **SimCLNMT (B)** is trained with candidate translations scored by BLEU, while SimCLNMT (C) is trained with candidate translations scored by COMET. The reported results represent the average scores across the test sets

### 5.2 Training with Different Evaluation Metrics

In our preceding experiments, BLEU is the primary metric for assessing the quality of candidate translations during the finetuning phase. To broaden the evaluation of our methodology beyond BLEU, we introduce a reference-based neural metric, COMET[6] (Rei et al., 2022), as the evaluation metric $M$ in Section 3.2. Subsequently, we train another version of **SimCLNMT**, leveraging the scores of candidate translations as calculated by COMET.

The results in Table 4 indicate that (1) Our model significantly enhances performance when BLEU or COMET is utilized as the target evaluation metric for ranking candidate translations. This suggests that our method can be optimized for any specific target metric, offering an alternative to reinforcement learning or minimum-risk training. (2) The model trained on one evaluation metric (e.g., COMET) also shows improvement on another metric (e.g., BLEU) compared with the MLE baseline model. This indicates that the enhancements from our model are not solely due to the potential weaknesses of individual metrics. Furthermore, it highlights a significant correlation between BLEU and COMET, reinforcing the generalization abilities of the model across different evaluation metrics.

### 5.3 Ranking Correlation

We evaluate the correlation between the model-calculated probabilities of candidate translations and their quality scores. We use Eq. 2 to estimate probabilities, using BLEU as the metric for the quality scores of the candidate translations. The Spearman rank correlation is calculated for each instance, with these scores then averaged to obtain the overall correlation. Using a total of 32 candidates for calculation, Table 5 shows that **SimCLNMT** exhibits a better rank correlation among candidate translations compared

[6] https://github.com/Unbabel/COMET

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1047-1058, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

to the MLE baseline model. This signifies that our model more accurately assigns higher probabilities to better candidate translations and lower probabilities to the less favorable ones.

| Model | English-Chinese | | | | |
|---|---|---|---|---|---|
| | NIST03 | NIST04 | NIST05 | NIST06 | average |
| MLE | 0.0812 | 0.0911 | 0.0583 | 0.0855 | 0.0790 |
| **SimCLNMT** | **0.1567** | **0.1449** | **0.1201** | **0.1240** | **0.1364** |

Table 5: Rank Correlation between the model's estimated probabilities of the candidate translations and the quality scores (BLEU) of the candidate translations on NIST (English-Chinese). We calculate the Spearman rank correlation between Transformer and **SimCLNMT**.

### 5.4 Statistical Significance Testing

In neural machine translation models, significance tests are crucial for evaluating the statistical significance of performance differences between model variations or different models. These tests help researchers determine whether observed performance improvements are due to random fluctuations or genuine enhancements in the model. We conduct statistical significance tests using the t-test to compare **SimCLNMT** with the MLE baseline model. The experimental results in Table 6 show that the p-value of the t-test is significantly less than 0.05, indicating a statistically significant difference. These findings further confirm that the performance of **SimCLNMT** is superior to that of the MLE baseline model.

| Dataset | NIST-ench | WMT14-ende |
|---|---|---|
| t-test | 0.0002842 | 0.001517 |

Table 6: The results of statistical significance testing

## 6 Limitation and Future Work

Table 7 offers a comprehensive statistical comparison between the MLE baseline and **SimCLNMT** models on the English-Chinese and English-German test sets. Specifically, for the NIST03 English-Chinese test set, both models generated 919 candidate translations, with their BLEU scores subsequently measured for evaluation. Upon analysis, in 33 cases, which represent 3.59% of the total instances, the BLEU scores for the **SimCLNMT** model were found to be lower than those of the MLE baseline. This percentage, termed the "decrease rate," indicates a decline in performance for the **SimCLNMT** model in certain sentences.

We hypothesize that this performance decline is attributable to the phenomenon of "forgetting," where the model loses some of the original knowledge post full-parameter fine-tuning. This insight has prompted us to consider strategies for future work aimed at enhancing translation quality and mitigating this forgetting effect. To address the issue, we plan to explore methods that include selectively freezing certain parameters within the MLE baseline model or incorporating adapter modules (Houlsby et al., 2019). These strategies are designed to preserve the model's previously acquired knowledge and prevent its loss, thereby potentially improving the stability and quality of translations produced by the **SimCLNMT** model.

| Dataset | English-Chinese | | | | English-German |
|---|---|---|---|---|---|
| | NIST03 | NIST04 | NIST05 | NIST06 | newstest2014 |
| decrease rate | 3.59% | 3.28% | 3.30% | 3.09% | 3.21% |

Table 7: The results of decrease rate.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1047-1058, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

## 7 Conclusion

In this study, we introduce a straightforward yet effective training method termed **SimCLNMT**. The objective of this approach is to select words that not only fit well in the immediate context but also enhance the overall translation quality over time. Through extensive experimentation on translation tasks between English-Chinese and English-German, we demonstrate that our **SimCLNMT** method consistently achieves substantial improvements. These results underscore the effectiveness of our approach in the domain of machine translation.

CCL 2024

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1047-1058, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

## References


Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018a. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 489–500.

Sergey Edunov, Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018b. Classical structured prediction losses for sequence to sequence learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 355–364.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898.

Markus Freitag and Yaser Al-Onaizan. 2017. Beam search strategies for neural machine translation. *arXiv preprint arXiv:1702.01806*.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *International conference on machine learning*, pages 1243–1252. PMLR.

Alex Graves. 2013. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*.

Evan Greensmith, Peter L Bartlett, and Jonathan Baxter. 2004. Variance reduction techniques for gradient estimates in reinforcement learning. *Journal of Machine Learning Research*, 5(9).

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1700–1709.

Julia Kreutzer, Joshua Uyheng, and Stefan Riezler. 2018. Reliability and learnability of human bandit feedback for sequence-to-sequence reinforcement learning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1777–1788.

Yixin Liu and Pengfei Liu. 2021. Simcls: A simple framework for contrastive learning of abstractive summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 1065–1072.

Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 2022. Brio: Bringing order to abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2890–2903.

Miguel Moura Ramos, Patrick Fernandes, António Farinhas, and André FT Martins. 2023. Aligning neural machine translation models: Human feedback in training and inference. *arXiv e-prints*, pages arXiv–2311.

Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Analyzing uncertainty in neural machine translation. In *International Conference on Machine Learning*, pages 3956–3965. PMLR.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Matt Post. 2018. A call for clarity in reporting bleu scores. *WMT 2018*, page 186.

Ricardo Rei, José GC De Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André FT Martins. 2022. Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. Minimum risk training for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1683–1692.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models (2016). *arXiv preprint arXiv:1610.02424*.

Zhen Yang, Wei Chen, Feng Wang, and Bo Xu. 2018. Improving neural machine translation with conditional sequence generative adversarial nets. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1346–1355.

Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. 2023. Rrhf: Rank responses to align language models with human feedback without tears. *arXiv preprint arXiv:2304.05302*.

CCL 2024

## Appendix

## A Training Details

During the pretraining phase, for the English-Chinese translation task, we adjust the learning rate to $5 \times 10^{-4}$, setting the warm-up steps at 4,000. We train the transformer-base model, as defined by Vaswani et al. (2017), for a total of 60,000 steps. We select the model checkpoint that achieves the best performance on the development dataset for further analysis. For the English-German translation task, we follow the configurations specified by Vaswani et al. (2017).

In the fine-tuning phase, we use FP16 to accelerate our training. The average training time for NIST and WMT14 is about 10 minutes. We use Adam optimizer with the learning rate set to $2 \times 10^{-4}$. The original development set is used for hyper-parameter optimization. Table 8 outlines the specific hyper-parameter configurations used during the fine-tuning phase.

| **Datasets** | $\alpha$ (Eq. 4) | $\lambda$ (Eq. 5) |
|---|---|---|
| NIST English-Chinese | 0 | 0.1 |
| WMT14 English-German | 1.5 | 1 |

Table 8: Hyper-parameter Setting

CCL 2024

# Translate-and-Revise: Boosting Large Language Models for Constrained Translation

**Pengcheng Huang**[1]*, **Yongyu Mu**[1]*, **Yuzhang Wu**[1], **Bei Li**[1], **Chunyang Xiao**[3], **Tong Xiao**[1,2]†, and **Jingbo Zhu**[1,2]
[1]NLP Lab, School of Computer Science and Engineering, Northeastern University, Shenyang, China
[2]NiuTrans Research, Shenyang, China
[3]JP Morgan, United Kingdom
hpc1449181552@outlook.com lixiaoyumu9@gmail.com
{xiaotong,zhujingbo}@mail.neu.edu.cn

## Abstract

Imposing constraints on machine translation systems presents a challenging issue because these systems are not trained to make use of constraints in generating adequate, fluent translations. In this paper, we leverage the capabilities of large language models (LLMs) for constrained translation, given that LLMs can easily adapt to this task by taking translation instructions and constraints as prompts. However, LLMs cannot always guarantee the adequacy of translation, and, in some cases, ignore the given constraints. This is in part because LLMs might be overly confident in their predictions, overriding the influence of the constraints. To overcome this overiding behaviour, we propose to add a revision process that encourages LLMs to correct the outputs by prompting them about the constraints that have not yet been met. We evaluate our approach on four constrained translation tasks, encompassing both lexical and structural constraints in multiple constraint domains. Experiments show 15% improvement in constraint-based translation accuracy over standard LLMs and the approach also significantly outperforms neural machine translation (NMT) state-of-the-art methods.

## 1 Introduction

Constrained translation seeks to generate translations that adhere to pre-specified constraints. To achieve this, conventional approaches impose constraints on machine translation systems and force them to follow the constraints during inference (Hokamp and Liu, 2017; Hasler et al., 2018; Dinu et al., 2019; Bergmanis and Pinnis, 2021b; Wang et al., 2022b; Ailem et al., 2022). More recently, large language models (LLMs) have been shown to be strong translation systems (Hendy et al., 2023; Moslem et al., 2023). They provide a general way to involve various instructions, demonstrations, and constraints into the translation process (Mu et al., 2023; Bogoychev and Chen, 2023), enabling us to perform constrained translation using off-the-shelf, well-trained LLMs.

While applying LLMs to constrained translation is straightforward, we observe empirically that even strong LLMs (i.e. `GPT-3.5`) do not always follow the instructions to obey constraints: LLMs' predictions often override the guide of constraints, which result in missing constraints during translation. See Figure 1 for an example where we use an LLM to translate an English sentence to a Chinese sentence with a lexical constraint "COVID-19→新型冠状病毒". We note that, despite significant effort in developing clear and instructive prompts, we were not able to improve the LLM in a single run of the LLM through the use of these constraints. For instance, we observed that when using open-source LLM to translate COVID-19, it tends to translate it as "新冠" more than 80% of the time, overlooking the constraint in the prompt to translate COVID-19 as "新型冠状病毒". The problem consists of a real use case for what describes as 'memo trap' in the LLM literature (McKenzie et al., 2023).

To alleviate this problem and thus improve the accuracy to meet constraints, we propose to construct prompts iteratively that enable better focus on the unsatisfied constraints. The idea behind our approach

*Equal contribution.
†Corresponding author.

©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1059-1074, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

is to leverage the auto correction skills of LLMs by explicitly prompting them with which constraints are not satisfied (Madaan et al., 2023; Zhang et al., 2023c; Jiang et al., 2023). To do this, we introduce a revision step after the initial run of LLMs where we provide the LLMs with both the already-generated translation and the constraints that have not been covered. Then, we instruct the LLM to revise its output by taking these constraints into account.

We conduct experiments across four diverse constrained translation datasets, encompassing two distinct constraint types: lexical and structural. Our proposed "Translate-and-Revise" (TAR) approach consistently elevates the performance of LLMs in constrained translation, achieving state-of-the-art (SoTA) results on multiple datasets.

The contributions of this work are as follows:

- We introduce a novel TAR strategy that initially employs LLMs as constraint-aware translators and subsequently reproposes them as revisers to revise translations that do not meet given constraints. We show that TAR significantly reduces missing constraints during translations.
- We rigorously evaluate our approach on four constrained translation datasets spanning multiple domains like news and electronics. Our results demonstrate a significant improvement in constraint fidelity and translation quality, outperforming existing methods and achieving SoTA results.
- To the best of our knowledge, our study is the first to evaluate LLMs across four distinct constrained translation datasets, thereby providing a robust LLM baseline for future research in the area. We believe our findings serve a solid baseline towards establishing more comprehensive benchmarks in the field of constrained translation.

## 2 Methods

Given a source language input and bilingual constraints, TAR first employs LLMs as translators for an initial translation. While this step often yields high-quality outputs, when the LLMs' confidence during generation exceeds the guidance of the constraints, it results in suboptimal translation outputs. To mitigate the occurrence of missing constraints in LLMs-based translation, we introduce a reviser to enhance adherence to the constraints in the translation. The revision process is iterated multiple times until all constraints are satisfied, or the maximum allowable number of modifications is reached. The process of TAR is provided in Figure 1. Next, we describe TAR in more details.

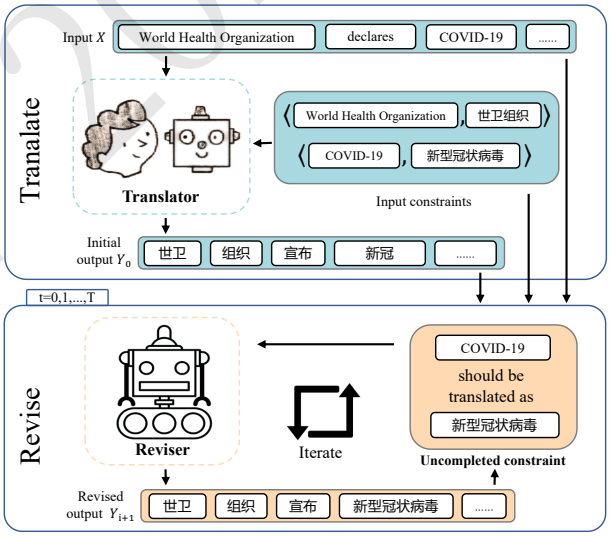



Figure 1: Given source language input $X$ and constraint pairs, a *Translator* produces an initial translation $Y_0$ where COVID-19 is translated as "新冠". Subsequently, a *Reviser* iteratively revise the translation $Y_i$ to a better one $Y_{i+1}$, correctly translating COVID-19 as "新型冠状病毒".

### 2.1 Translate

Let $X = \{x_1, x_2, ..., x_n\}$ be the source-language sentence with length $n$, and $Y = \{y_1, y_2, ..., y_m\}$ be the target-language sentence with length $m$. The translation procedure can be written as:

$$Y = \text{Trans}(f(X)) \tag{1}$$

where $\text{Trans}(\cdot)$ symbolizes the translation model (either an NMT model or an LLM), and $f(\cdot)$ denotes a template by which we process $X$ to make it suitable as the input of $\text{Trans}(\cdot)$.

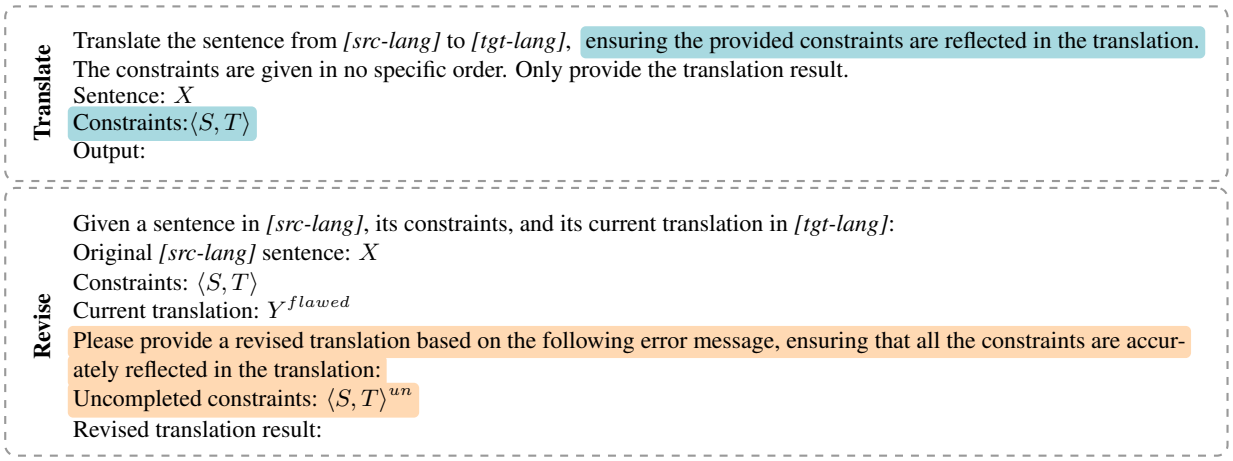



Figure 2: Two stages of TAR, in the Translate stage, constraints $\langle S, T \rangle$ are incorporated into the prompt to enable the model to generate preliminary translation results that meet the constraints to a certain extent. In the Revise stage, LLMs revise the flawed translation results $Y^{flawed}$ with uncompleted constraints $\langle S, T \rangle^{un}$. The sections shaded in blue and yellow respectively represent the important parts of the two stages.

Let the $\langle S, T \rangle = \{\langle s_1, t_1 \rangle, \langle s_2, t_2 \rangle, ..., \langle s_k, t_k \rangle\}$ represents the bilingual constraints with $k$ pairs in total. Constrained translation needs the system to accurately translate each source constraint $s_i$ to its corresponding target constraint $t_i$. This process can be represented by the following equation:

$$\begin{aligned} Y &= \text{Trans}(f(X, \langle S, T \rangle)) \\ s.t.\ & \ T \in Y \end{aligned} \tag{2}$$

Since conventional translation instructions never impose constraints on LLMs, they frequently fall short of satisfying constraints. In this work, we propose to integrate these constraints directly into the prompts and employ an instruction based on natural language specifically tailored for constrained translation tasks. Our template $f(\cdot)$ is shown in Figure 2 which can effectively turn LLMs into constraint-aware translators.

## 2.2 Revise

However, the LLM-based translation cannot always cover all original constraints. We randomly sampled 20 incorrect translation results and observed that, in datasets like WMT21 Terminology Translation (Alam et al., 2021), to $95\%(19/20)$ of the cases, the tokens generated by the model were similar to the expected constraints meaning and exhibited high confidence levels. The confidence level of LLMs in generating these tokens remained virtually unchanged, whether or not constraints were included in the instructions, revealing overconfidence in generation while overlooking the constraints.

We notice a strong connection between our real use case and 'memo trap' (McKenzie et al., 2023) as unsatisfied constraints often pertain to non-mainstream translations resulting terms used with lower frequency and the incorrect translations usually refer to the mainstream translations. Compared 'memo trap', we show that the phenomenon extends to toy settings and is prominent even for real applications and for SOTA models like GPT-3.5.

To overcome these challenges, we initially employ a rule-based method to identify which constraints are not completed. Subsequently, these uncompleted constraints $\langle S, T \rangle^{un}$, along with the source language input $X$, flawed translations output $Y^{flawed}$, and all other given constraints $\langle S, T \rangle$, are passed to the LLM. At this juncture, the LLM assumes the role of a reviewer, tasked with revising flawed translation upon receipt of uncompleted constraints. The aforementioned process is defined by the following formula:

$$Y = \text{Revise}(f(X, \langle S, T \rangle, \langle S, T \rangle^{un}, Y^{flawed})) \tag{3}$$

CCL 2024

where $\mathrm{Revise}(\cdot)$ symbolizes the reviser. Furthermore, TAR can continuously iterate in the loop of detecting uncompleted constraints and making revisions until a stopping condition is met. This condition is either the iteration $i$ reaches a specified count or the translation satisfies all constraints. We represent this iterative process as follows:

$$Y_{i+1} = \mathrm{Revise}(f(X, \langle S, T \rangle, \langle S, T \rangle_i^{un}, Y_i)) \tag{4}$$

where the translation $Y_i$ from the previous iteration, combined with its uncompleted constraints $\langle S, T \rangle_i^{un}$ and other inputs, is sent into the reviser to produce a more precise translation $Y_{i+1}$. By highlighting uncompleted translation constraints and comparing flawed translation results, human translators are able to satisfy these constraints and optimize translation output. Empirically, we find that LLMs can revise the translation results similarly to human translators, while being more efficient and cost-effective. Additionally, we discuss the time and financial costs of multiple iterations in Appendix A.

## 3 Experiments

In this study, we evaluate the performance of the TAR in constrained translation. While most previous research has typically focused on just one or two constrained translation tasks (Dinu et al., 2019; Wang et al., 2022a; Hashimoto et al., 2019; Zeng et al., 2023), our evaluation expands to two types of constraints: lexical constraints and structural constraints, covering four practical scenarios: general lexically constrained translation, translation with terminology constraints, translation with named entity constraints, and structured document translation.

### 3.1 Setup

**Datasets** Detailed information about the datasets we used can be found in Table 1. Lexical constraints refer to sentences with predefined word or phrase constraints sourced from existing databases. Structural constraints, in contrast, encompass inline markup tag constraints like XML tags, for example, <ph> and </ph>. Here are more details about the datasets that we use in this work:

*General Lexically Constrained Translation:* This is based on a dataset[0] provided by (Dinu et al., 2019). The dataset is derived from *newstest2017* En → De. Lexical constraints are extracted with guidance from two general-domain term databases: IATE and Wiktionary[1].

*Terminology Translation:* To benchmark against SoTA NMT systems, we employ the official test set from the *terminology translation* task in WMT21[2](WMT21 TT) (Alam et al., 2021).

| Corpus | Language direction | #Sent. | #Lines with const. | #Const. |
|---|---|---|---|---|
| *Lexical constraint* | | | | |
| IATE | En-De | 2000 | 414 | 452 |
| Wiktionary | En-De | 2000 | 727 | 884 |
| WMT21 TT | En-Ru | 2100 | 1307 | 2524 |
| | En-Zh | 2100 | 1191 | 2229 |
| ETC | En-Zh | 19144 | 12040 | 35253 |
| | En-Ru | 12985 | 3917 | 10308 |
| | Zh-En | 19144 | 12040 | 35253 |
| | Ru-En | 12985 | 3917 | 10308 |
| *Structural constraint* | | | | |
| LXM | En-Zh | 2000 | 518 | 884 |
| | En-De | 2000 | 520 | 942 |
| | En-Ru | 2000 | 554 | 993 |
| | En-Fr | 2000 | 575 | 1051 |

Table 1: Statistics of the datasets we used for four different task: the total number of sentences in datasets (*#Sent.*), the number of lines with constraints (*#Lines with const.*), the number of constraints (*#const.*).

*Entity Translation:* We also endeavor to evaluate our method using the extensive Entity Translation Corpus (ETC) (Zeng et al., 2023), which comprises six test sets from the WMT News Translation Task spanning 2015-2021. For alignment, we employ spaCy NER models[3] to extract source entities and use awesome-align (Dou and Neubig, 2021) for their correspondence.

*Structured Document Translation:* Following recent works (Wang et al., 2022a), we conduct experiments on the LXM dataset[4] (Hashimoto et al., 2019), in which XML tags are hierarchically distributed

[0] https://github.com/mtresearcher/terminology_dataset
[1] Available at https://iate.europa.eu/home and https://www.wiktionary.org/.
[2] https://www.statmt.org/wmt21/terminology-task.html
[3] https://pypi.org/project/spacy/
[4] https://github.com/salesforce/localization-xml-mt

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1059-1074, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

| Method | BLEU | CCR% | BLEU | CCR% |
|---|---|---|---|---|
| *Dataset* | *IATE* | | *Wiktionary* | |
| Transformer | 25.8 | 76.3 | 26.0 | 76.9 |
| Const. Dec. | 25.3 | 82.0 | 25.8 | **99.5** |
| Code-switching | 26.0 | 94.5 | 26.3 | 93.4 |
| Append | 26.0 | 92.8 | 26.9 | 90.7 |
| RTT | 27.2 | **99.6** | 27.8 | 98.3 |
| LLM Trans. | **32.0** | 85.4 | *32.0* | 88.7 |
| LLM Const. Trans. | **32.0** | 96.2 | **32.1** | 97.2 |
| +Revision | **32.0**$_{(+0.0)}$ | *98.9*$_{(+2.7)}$ | *32.0*$_{(-0.1)}$ | *98.9*$_{(+1.7)}$ |

Table 2: Results of the general lexically constrained translation task. The **highest** scores among the various systems are highlighted in bold, while the *second-best* scores are emphasized in italics for clarity.

| Method | BLEU | CCR% | BLEU | CCR% |
|---|---|---|---|---|
| *Direction* | *English-Chinese* | | *English-Russian* | |
| HW-TSC | **40.7** | 88.6 | - | - |
| TermMind-sys2 | *40.5* | 85.6 | - | - |
| ProMT.soft | - | - | **31.1** | 90.9 |
| TildeMT | - | - | 28.2 | 86.3 |
| Lingua Custodia | 29.6 | 82.8 | 28.8 | 85.4 |
| LLM Trans. | 36.3 | 87.2 | 29.7 | 85.9 |
| LLM Const. Trans. | 36.4 | *92.6* | 30.1 | *95.8* |
| +Revision | 35.9$_{(-0.5)}$ | **95.9**$_{(+3.3)}$ | *30.3*$_{(+0.2)}$ | **97.5**$_{(+1.7)}$ |

Table 3: Results of the terminology translation task for both English-Chinese and English-Russian.

throughout the source and target text.

**Evaluation Metrics** Consistent with previous studies (Dinu et al., 2019; Ailem et al., 2022; Wang et al., 2022a; Zeng et al., 2023), we employ BLEU (Papineni et al., 2002) and the constraints completion rate (CCR) to assess translation quality and constraint-based translation accuracy except for structured document translation. For entity translation, we also incorporate the COMET score[5] (Rei et al., 2020) for comparison with the work of Zeng (2023). In structured document translation, we utilize sacreBLEU (Post, 2018) to compute the XML-based BLEU score[6]. Additionally, we measure structured constraint-based translation accuracy using the structure accuracy rate (SAR) and structure match rate (SMR). Here, SAR evaluates the compatibility of translation results with XML parsers, while SMR ensures the translated XML structure aligns with the reference. Both metrics are assessed using lxml[7].

**Baselines** To ensure thorough evaluation, apart from comparing TAR with the LLM baseline without revision, we also we compare TAR with representative methods across each of the four tasks. For general lexically constrained translation, our baselines include the vanilla Transformer (Vaswani et al., 2017), Const.Dec. (Post and Vilar, 2018), Code-switching (Dinu et al., 2019), Append (Dinu et al., 2019), and Robust Terminology Translation (RTT) (Zhang et al., 2023b). For the terminology translation task, our baselines are derived from the top three submissions of WMT21. They include HW-TSC (Wang et al., 2021b), Term-Mind-sys2 (Wang et al., 2021a), ProMT.soft (Molchanov et al., 2021), TildeMT (Bergmanis and Pinnis, 2021a), and Lingua Custodia (Ailem et al., 2021). For entity translation, we utilize the vanilla Transformer, Code-switching, Placeholder (Yan et al., 2019), and Extract and Attend (Zeng et al., 2023). Structured document translation baselines consist of the vanilla Transformer, Split-Inject (Al-Anzi et al., 1997), and Template (Wang et al., 2022a) methods.

**Model Configurations** We initially investigate the potential of LLMs to act as both the translator and reviser within TAR. Our primary choice for LLMs is `gpt-3.5-turbo-0613`[8], chosen for its exceptional translation capabilities and proficiency in adhering to instructions. Additionally, we assess its revising process for NMT models in Section 4.1 and explore whether TAR provides consistent improvements across different LLMs in Section 4.2. The decoding parameters for these models remain at their default settings, except for the sampling temperature, which is set to 0. We employ natural language-based prompts in a one-shot manner, merging uncompleted constraints with the source language, flawed translation results, and original constraints to form the reviser's input. These prompts are depicted in Figure 2.

**Detection of Uncompleted Constraints** To identify unmet constraints in translations, we employ a rule-based procedure that leverages scripts designed for calculating CCR. This procedure assesses how well the translation adheres to the constraints. We further explore the capacity of LLMs to autonomously verify constraint completion and offer detailed feedback to the reviser in Section 3.3.

[5] `wmt22-comet-da`

[6] XML tags are treated as an integral part of the sentences during BLEU score calculation.

[7] https://lxml.de/

[8] `https://openai.com/`

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1059-1074, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

| Method | BLEU | COMET | CCR% | BLEU | COMET | CCR% |
|---|---|---|---|---|---|---|
| *Direction* | *English-Chinese* | | | *Chinese-English* | | |
| Transformer | 26.3 | 34.8 | 57.3 | 27.5 | 41.5 | 59.0 |
| Code-switching | 25.9 | 41.4 | 70.5 | 27.2 | 45.0 | 71.1 |
| Placeholder | 26.4 | 42.9 | 71.4 | 27.5 | 47.2 | 72.1 |
| Extract & Attend | 26.8 | 48.6 | 72.3 | *28.0* | 50.1 | 72.5 |
| LLM Trans. | 39.5 | **87.6** | 77.2 | 27.3 | **83.8** | 85.3 |
| LLM Const. Trans. | **40.0** | *87.4* | *96.6* | **28.9** | *83.5* | *94.2* |
| +Revision | **40.0**$_{(+0.0)}$ | *87.4*$_{(+0.0)}$ | **97.6**$_{(+1.0)}$ | **28.9**$_{(+0.0)}$ | *83.5*$_{(+0.0)}$ | **97.3**$_{(+3.1)}$ |
| *Direction* | *English-Russian* | | | *Russian-English* | | |
| Transformer | 31.8 | 52.2 | 40.0 | 34.6 | 54.0 | 48.7 |
| Code-switching | 30.5 | 55.2 | 50.4 | 32.0 | 56.7 | 50.2 |
| Placeholder | 31.9 | 57.6 | 50.3 | 34.7 | 59.1 | 50.7 |
| Extract & Attend | **32.7** | 62.2 | 57.3 | 35.4 | 63.5 | 58.4 |
| LLM Trans. | 31.8 | **89.9** | 64.5 | *36.0* | **85.8** | 76.8 |
| LLM Const. Trans. | *32.6* | *89.8* | *88.8* | **36.8** | **85.8** | *96.7* |
| +Revision | 32.5$_{(-0.1)}$ | *89.8*$_{(+0.0)}$ | **89.8**$_{(+1.0)}$ | **36.8**$_{(+0.0)}$ | **85.8**$_{(+0.0)}$ | **97.5**$_{(+0.8)}$ |

Table 4: Results of the entity translation task for English-Chinese, English-Russian, Chinese-English and Russian-English.

| Method | BLEU | SAR% | SMR% | BLEU | SAR% | SMR% |
|---|---|---|---|---|---|---|
| *Direction* | *English-Chinese* | | | *English-German* | | |
| Transformer | *61.2* | 99.85 | 99.25 | *52.7* | 99.80 | 99.20 |
| Split-Inject | 57.0 | **100.00** | 99.30 | 50.7 | **100.00** | **99.80** |
| Template | **61.5** | **100.00** | **99.80** | **53.6** | **100.00** | **99.80** |
| LLM Trans. | 55.1 | *99.95* | 98.95 | 49.2 | *99.95* | 99.25 |
| LLM Const. Trans. | 56.4 | **100.00** | 99.50 | 49.2 | *99.95* | 99.25 |
| +Revision | 56.5$_{(+0.1)}$ | **100.00**$_{(+0.00)}$ | *99.75*$_{(+0.25)}$ | 49.2$_{(+0.0)}$ | **100.00**$_{(+0.05)}$ | 99.30$_{(+0.05)}$ |
| *Direction* | *English-French* | | | *English-Russian* | | |
| Transformer | *65.3* | 99.55 | 99.30 | *44.9* | 99.45 | 98.90 |
| Split-Inject | 66.1 | **100.00** | **100.00** | 43.1 | **100.00** | 99.85 |
| Template | **67.3** | **100.00** | **100.00** | **45.8** | **100.00** | **99.80** |
| LLM Trans. | 58.1 | *99.90* | 99.30 | 34.4 | *99.90* | 99.35 |
| LLM Const. Trans. | 59.3 | **100.00** | *99.95* | 36.0 | **100.00** | 99.60 |
| +Revision | 59.3$_{(+0.0)}$ | **100.00**$_{(+0.00)}$ | *99.95*$_{(+0.00)}$ | 36.0$_{(+0.0)}$ | **100.00**$_{(+0.00)}$ | *99.75*$_{(+0.15)}$ |

Table 5: Results of the structured document translation for English-Chinese, English-German, English-Chinese and English-Russian.

### 3.2 Main results

Table 2, Table 3, Table 4, and Table 5 detail the performance of TAR on general lexically constrained translation, terminology translation, entity translation, and structured document translation, respectively. Here are our main results.

**Comparison with base LLMs** TAR consistently boosts the performance of LLMs in constrained translation. Two primary factors contribute to this improvement:

(1) Our natural language-based prompts, as opposed to the conventional few-shot translation prompts (Hendy et al., 2023), are more effective for constrained translation. Specifically, in terminology translation (refer to Table 3), our prompts lead to an average BLEU score increase of 0.3 and a CCR rise of 7.7%.

We noticed significant gains over base LLMs in entity translation. Across four language directions, there is a consistent uplift in BLEU scores, averaging an increase of 0.9. Notably, the CCR experiences increases of 19.4%, 8.9%, 24.3%, and 19.9% respectively. This marked improvement can primarily be attributed to the superior instruction-following capabilities of LLMs. By incorporating constraints in instructions, we can guide the model more effectively to address these constraints, alleviating the issue of LLMs struggling to correctly translate named entities.

CCL 2024

(2) Revision effectively improves constraint-based translation accuracy across all datasets without sacrificing translation quality, addressing the tendency of LLM-based translations to overlook constraints.

As illustrated in Table 6, we observe that the LLM exhibits overconfidence in its prediction overriding the influence of constraints, especially when the constraint suggests an uncommon translation for a polysemous word. For instance, the LLM translates "COVID-19" to the more commonly used "新冠" instead of adhering to the target constraint "新型冠状病毒". We speculate this overconfidence in LLMs stems from their greater exposure to "新冠" compared to "新型冠状病毒" during the pre-training phase, which may lead them to be overly loyal to certain patterns, thereby preventing them from meeting certain constraints.

| | |
|---|---|
| **Constraints** | ⟨WHO, 世卫组织⟩; ⟨COVID-19, 新型冠状病毒⟩ |
| **Source** | On 11 March 2020, WHO characterized COVID-19 as a pandemic. |
| **Reference** | 2020年3月11日 世卫组织 将 新型冠状病毒 列为大流行病。 |
| **Const. Trans.** | 2020年3月11日 世卫组织 将新冠确定为大流行病。 |
| **+ Revision** | 2020年3月11日 世卫组织 将 新型冠状病毒 定性 为大流行病。 |

Table 6: A case study of TAR: Initially, the translator rendered "COVID-19" as the more prevalent "新冠" in Chinese. With the intervention of the reviser, it was accurately translated as "新型冠状病毒", thereby satisfying all constraints.

However, the revision step can effectively cut down the possibility of missing constraints by explicitly prompting LLMs with which constraints are not satisfied. Experimental results indicate that our revision strategy led to average improvements in CCR by 2.2% for lexically constrained translation, 2.5% for terminology translation, and 1.5% for entity translation across various language directions. While the improvements in SCR and SMR for structured document translation might not seem prominent, it's primarily because the initial translation is already at a 100% performance.

**Comparison with supervised methods** The SoTA methods for these four constrained translation datasets predominantly rely on pseudo-data augmentation. Through our experimental results, we observe that these methods nearly achieve perfection on the IATE, Wiktionary, and LXM datasets (as evidenced in Table 2 and Table 5). We contend that the test sets for these datasets might be relatively straightforward, and the constraints they encompass are frequently encountered in training sets. Therefore, they may not accurately reflect real-world applications where constraints might span multiple domains and are infrequently seen in the training data. However, when assessed on the ETC (as depicted in Table 4), which comprises test data spanning from 2015 to 2021, showcasing a rich diversity in constraint domains, the efficacy of traditional data augmentation methods seems to be poor. In contrast, TAR's performance remains stable, demonstrating comparable constraint-based translation accuracy on ETC as with other datasets. This highlights TAR's proficiency in handling the diverse constraint requirements found in real-world situations.

### 3.3 Impact of Inputs on Reviser Performance

The reviser receives inputs including the source language sentence, translation results, given constraints, and the uncompleted constraints. To evaluate the significance of each component, we conducted experiments wherein we omitted specific elements from the input. The variations include: 1) Excluding uncompleted constraints; 2) Excluding original constraints; and 3) Only indicating to the model that the translation is flawed without specifying the uncompleted constraints. All other settings remain unchanged. The comparative outcomes on IATE and Wiktionary are presented in Table 7.

| | IATE | | Wiktionary | |
|---|---|---|---|---|
| Setting | BLEU | CCR% | BLEU | CCR% |
| base | 32.0 | 96.2 | 32.1 | 97.2 |
| after revise | 32.0 | **98.9** | 32.1 | **98.9** |
| - Uncompleted const. | 32.0 | 97.4 | 32.1 | 97.9 |
| - Original const. | 32.0 | 97.1 | 32.0 | 97.7 |
| - Both | 32.0 | 96.2 | 32.1 | 97.2 |
| + Detected by LLM | 31.9 | 95.8 | 32.0 | 97.2 |

Table 7: BLEU and CCR scores of ablation on supplementary feedback. "Uncompleted constraints" and "Original constraints" are parts of the input received by the reviser.

From our observations, the CCR scores of the variants show a decline compared to the default input of the TAR reviser on both datasets. Interestingly, the omission of the original constraints has a more pronounced impact on CCR. This could be attributed to the negative modification of completed constraints made by the reviser when it is inaccessible to the given constraints. Additionally, when both elements

are excluded, there's no noticeable difference in the CCR before and after the revision process.

Furthermore, as shown by '+ Detected by LLM' in Table 7, using LLMs to detect uncompleted constraints and then feeding them back to the reviser may degrade translation performance. Further analysis of the results reveals challenges faced by LLMs inaccurately identifying unsatisfied constraints; they often mistakenly believe certain constraints have been met. Such inaccurate feedback not only fails to enhance the quality of the translation but might even deteriorate it. These insights emphasize the criticality of supplying the reviser with exact and thorough constraint information.

## 4 Analysis

### 4.1 TAR Augments NMT Translators

In our study, we initially depended on constraint-aware translators to produce preliminary translation results. However, in real-world scenarios, industry practitioners often possess powerful domain-agnostic NMT models. These models, due to their lack of training with specific constraints, frequently fall short in constrained translation tasks. In this section, we integrate TAR into these general-purpose NMT models. By iteratively optimizing the NMT translation results through TAR, we can significantly enhance the CCR of the translation while ensuring its quality.

| | BLEU | CCR% | BLEU | CCR% |
|---|---|---|---|---|
| **WMT21 TT** | *English-Chinese* | | *English-Russian* | |
| TAR | 35.9 | 95.9 | 30.3 | 97.5 |
| NMT | 34.5 | 85.6 | 33.6 | 85.3 |
| + Revision | 35.2$_{(+0.7)}$ | 95.6$_{(+10.0)}$ | 34.0$_{(+0.4)}$ | 96.6$_{(11.3)}$ |
| **ETC** | *English-Chinese* | | *English-Russian* | |
| TAR | 40.0 | 97.6 | 32.5 | 89.8 |
| NMT | 41.2 | 74.5 | 40.1 | 72.3 |
| + Revision | 42.7$_{(+1.5)}$ | 92.8$_{(+18.3)}$ | 40.2$_{(+0.1)}$ | 82.4$_{(+10.1)}$ |

Table 8: Results of applying TAR to NMT on the WMT21 TT and ETC datasets for both English-Chinese and English-Russian.

Specifically, we first the WMT21 champion model (Tran et al., 2021) to obtain a preliminary translation result. Since this model is not specifically trained for constraints, the initial translation often exhibits a suboptimal CCR. Building on this, we apply TAR to revise this outcome, iteratively optimizing to form the final translation result.

The experimental results on WMT21 TT and ETC datasets are presented in Table 8, we can see that TAR bolsters both BLEU and CCR scores of NMT models. On average, we observed an uplift of 0.7 and 0.8 in BLEU scores, coupled with impressive gains of 10.7% and 14.3% in CCR across the two datasets and language pairs. Although there remains a gap in constraint-based translation accuracy compared to the standard TAR, it generally exhibits superior translation quality. This insight demonstrates that when equipped with TAR, even domain-agnostic NMT models can adeptly tackle constrained translation. This eliminates the need for forced decoding algorithms or additional training, greatly enhancing their usability in constrained translation applications.

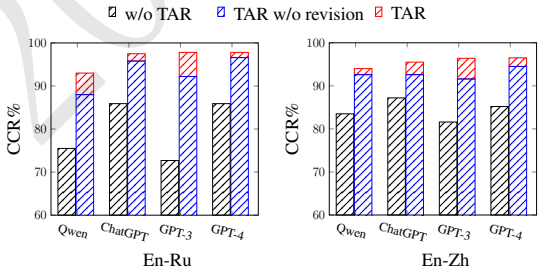



Figure 3: TAR results on WMT21 TT using `Qwen`, `ChatGPT`, `GPT-3 (text-davinci-003)` and `GPT-4`. Here "w/o TAR" represents the use of the conventional translation prompt. "TAR w/o revision" indicates the use of a prompt with constraints, but without reviser. Meanwhile, "TAR" denotes the full method that includes revisions.

### 4.2 Scaling TAR to More LLMs

To evaluate the scalability and robustness of TAR across different models, we applied it to a variety of LLMs, including commercial models like `GPT-3` and `GPT-4`, as well as the open-source `Qwen`[9] (Bai et al., 2023). We maintained consistency in all other settings. Experiments were conducted in the "En-Zh" and "En-Ru" language directions for terminology translation. As shown in Figure 3, TAR consistently improves the performance of various LLMs in constrained translation tasks. To be specific, the average CCR for `GPT-3`, `GPT-4`, and `Qwen` increases by 20%, 11.7%, and 14.4%, respectively. Comparing the results of `ChatGPT` with those of `GPT-3` and `GPT-4`, it's evident that TAR enables more powerful models to fully harness their capabilities in constrained translation. Intriguingly, although the CCR score

[9]Qwen-14B-Chat

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1059-1074, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

of GPT-3 in the initial translation substantially trails that of ChatGPT, it surpasses ChatGPT post-revision. While the performance of Qwen lags slightly behind ChatGPT, the improvement brought by TAR is still notable.

### 4.3 Revision Iterative Round and Prompt Ensemble

The reviser, in its function, takes the translation result and uncompleted constraints as input. Naturally, one might consider iteratively revising the output multiple times. The question arises: how many iterations strike the optimal balance? Here, we employed constraint-aware LLMs and NMT on the "En-Zh" and "En-Ru" directions of the terminology translation dataset. We assessed performance across different iterative rounds of the revision module, consistently using the same prompts for each iterative phase.

As presented in Figure 4a, there is a significant leap in performance predominantly during the initial revision. Although performance does enhance with increasing iterations, the rate of improvement starts to taper off, indicating diminishing returns.

Furthermore, we also investigated whether the LLM reviser can benefit from varying prompts in multiple iterations. To experiment this, after designing several revision templates, we randomly select one for each iterative round. Thus, a complete multi-round revision process can utilize various templates. This design is in part similar to prompt ensemble methods (Zhang et al., 2023a; Pitis et al., 2023), combining benefits of various prompts, akin to ensemble learning (Dong et al., 2020). Experimental results are shown in Figure 4b. Compared to applying a single template in revision iterations, utilizing diverse templates achieves superior performance.

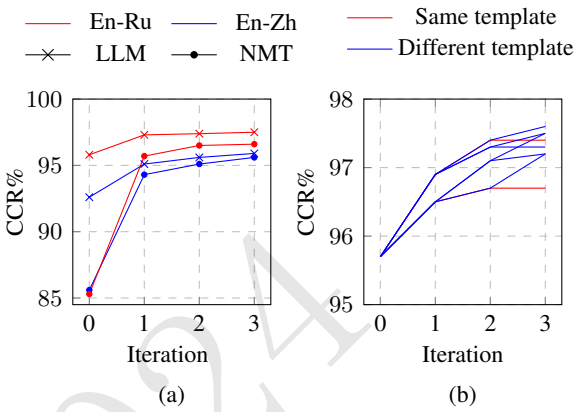



Figure 4: **(a)** Improvements in CCR with each iteration. **(b)** Red denotes consistent template use across three iterations, while blue indicates alternating templates.

### 4.4 More Analysis

Due to space limitations, we provide a more detailed analysis of our method in the appendix, including the additional costs incurred by TAR, the impact on performance as the number of constraints increases, reasons for potential declines in BLEU scores during the revision phase for certain datasets, and the performance of traditional constrained translation data augmentation methods on LLMs.

## 5 Related Work

### 5.1 Constrained Translation

Machine translation has made considerable progress in incorporating pre-specified constraint, which can be categorized into hard constrained translation and soft constrained translation.

**Hard constrained translation:** This line of research expanded the original search space via decoding algorithm modification to strictly incorporate constraints (Hokamp and Liu, 2017; Post and Vilar, 2018; Hu et al., 2019). However, while these methods achieve a high constraint-based translation accuracy, they tend to be computationally expensive and can sometimes compromise translation quality (Hasler et al., 2018; Zhang et al., 2021).

**Soft constrained translation:** Here, research primarily centers on data augmentation strategies to train NMT models to integrate constraints. Several techniques have been proposed, including: replacing source language constraints with special token (Crego et al., 2016; Wang et al., 2017; Zhang et al., 2023b); substituting source language constraints with target language constraints (Song et al., 2019; Dinu et al., 2019); using inline annotations to individually mark source and target language constraints (Ailem

CCL 2024

et al., 2022; Chen et al., 2021b); and (Wang et al., 2022a) employing a template to transform the constrained translation into constraint reordering. There are also studies that modified the model architecture to better integrate vectorized constraints representation (Li et al., 2020; Wang et al., 2022b), alignment information (Chen et al., 2021a). These approaches heavily rely on data quality or necessitate structural modifications, limiting their practicality. Moreover, they often falter when addressing diverse real-world requirements. In contrast, TAR doesn't require training on constraint-specific data and is adept at handling varied constraint scenarios.

### 5.2 Automatic Post-editing

Several studies have aimed to develop neural-based models for automatic post-editing (APE) in translation (Vu and Haffari, 2018; Correia and Martins, 2019; Shterionov et al., 2020). Chatterjee(2019) investigated the application of deep learning techniques for APE and introduced novel architectures to improve the quality of post-edited translations. Góis(2020) examined the application of automated ordering methods to improve translations. Voita(2019) introduced a context-aware approach to APE, integrating source context information into the neural framework to produce improved post-edits. Chollampatt(2020) investigated the application of APE in enhancing the translation performance of NMT models. These methods primarily focus on enhancing the overall translation quality. However, it's crucial to understand that not all words within a sentence carry equal importance. The precise translation of terminologies and entities significantly impacts user experience. Our proposed TAR specifically addresses the challenge of ensuring more accurate translations for these constraints.

## 6 Conclusion

In this work, we introduce the TAR prompting method, adeptly leverages LLMs for constrained translation. Our approach involves a two-step process: first using LLMs for constrained translation, and subsequently deploying them to revise translations with uncompleted constraints. Our approach mainly improves the constraint accuracy while maintaining translation quality by overcoming the 'memo trap' from the LLMs during translation using dedicated revision prompts in an iterative manner.

We further show that TAR can be applied to LLM based translation systems as well as traditional NMT systems, in both cases resulting in better constraint accuracy while maintaining translation quality and the technology is not limited to particular LLMs. More generally, our study sheds light on the importance of accurate feedback in general for LLM revision to work effectively.

## 7 Limitations

While we have demonstrated TAR's efficacy across four constrained translation datasets, real-world applications are considerably more varied, our prompts might not always yield optimal outcomes. In fact, the essence of TAR lies in its revision mechanism. However, as emphasized in Section 3.3, detecting constraint adherence using LLMs poses challenges. Rule-based methods, though effective in offering accurate feedback to the reviser, can falter in broader constraint scenarios, such as controlled text generation demanding specific stylistic alignment. In such contexts, devising a method to secure accurate and efficient feedback to guide model revisions remains a research imperative. We believe that overcoming these challenges will solidify TAR's standing as a universally effective framework across diverse constraint scenarios.

## Acknowledgements

This work was supported in part by the National Science Foundation of China (No.62276056), the Natural Science Foundation of Liaoning Province of China (2022-KF-16-01), the Fundamental Research Funds for the Central Universities (Nos. N2216016 and N2316002), the Yunnan Fundamental Research Projects (No. 202401BC070021), and the Program of Introducing Talents of Discipline to Universities, Plan 111 (No.B16009).

## References


Melissa Ailem, Jinghsu Liu, and Raheel Qader. 2021. Lingua custodia's participation at the wmt 2021 machine translation using terminologies shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 799–803.

Melissa Ailem, Jinghsu Liu, and Raheel Qader. 2022. Encouraging neural machine translation to satisfy terminology constraints. In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, page 446.

F Al-Anzi, K Al-Zame, M Husain, and H Al-Mutairi. 1997. Automatic english/arabic html home page translation tool. In *Proc. 1st Workshop Technol. Arabizing Internet*.

Md Mahfuz Ibn Alam, Ivana Kvapilíková, Antonios Anastasopoulos, Laurent Besacier, Georgiana Dinu, Marcello Federico, Matthias Gallé, Kweon Woo Jung, Philipp Koehn, and Vassilina Nikoulina. 2021. Findings of the WMT shared task on machine translation using terminologies. In Loïc Barrault, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno-Yepes, Philipp Koehn, Tom Kocmi, André Martins, Makoto Morishita, and Christof Monz, editors, *Proceedings of the Sixth Conference on Machine Translation, WMT@EMNLP 2021, Online Event, November 10-11, 2021*, pages 652–663. Association for Computational Linguistics.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report.

Toms Bergmanis and Mārcis Pinnis. 2021a. Dynamic terminology integration for covid-19 and other emerging domains. In *Proceedings of the Sixth Conference on Machine Translation*, pages 821–827.

Toms Bergmanis and Mārcis Pinnis. 2021b. Facilitating terminology translation with target lemma annotations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3105–3111.

Nikolay Bogoychev and Pinzhen Chen. 2023. Terminology-aware translation with constrained decoding and large language model prompting. *arXiv preprint arXiv:2310.05824*.

Rajen Chatterjee. 2019. Automatic post-editing for machine translation. *arXiv preprint arXiv:1910.08592*.

Guanhua Chen, Yun Chen, and Victor OK Li. 2021a. Lexically constrained neural machine translation with explicit alignment guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12630–12638.

Guanhua Chen, Yun Chen, Yong Wang, and Victor OK Li. 2021b. Lexical-constraint-aware neural machine translation via data augmentation. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3587–3593.

Shamil Chollampatt, Raymond Hendy Susanto, Liling Tan, and Ewa Szymanska. 2020. Can automatic post-editing improve nmt? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2736–2746.

Gonçalo M Correia and André FT Martins. 2019. A simple and effective approach to automatic post-editing with transfer learning. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 3050–3056.

Josep Crego, Jungi Kim, Guillaume Klein, Anabel Rebollo, Kathy Yang, Jean Senellart, Egor Akhanov, Patrice Brunelle, Aurelien Coquard, Yongchao Deng, et al. 2016. Systran's pure neural machine translation systems. *arXiv preprint arXiv:1610.05540*.

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 3063–3068.

Xibin Dong, Zhiwen Yu, Wenming Cao, Yifan Shi, and Qianli Ma. 2020. A survey on ensemble learning. *Frontiers Comput. Sci.*, 14(2):241–258.


Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1059-1074, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128.

António Góis, Kyunghyun Cho, and André Martins. 2020. Learning non-monotonic automatic post-editing of translations from human orderings. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 205–214.

Kazuma Hashimoto, Raffaella Buschiazzo, James Bradbury, Teresa Marshall, Richard Socher, and Caiming Xiong. 2019. A high-quality multilingual dataset for structured documentation translation. In *Proceedings of the Fourth Conference on Machine Translation*, pages 116–127.

Eva Hasler, Adrià De Gispert, Gonzalo Iglesias, and Bill Byrne. 2018. Neural machine translation decoding with terminology constraints. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 506–512.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Chris Hokamp and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1535–1546.

J Edward Hu, Huda Khayrallah, Ryan Culkin, Patrick Xia, Tongfei Chen, Matt Post, and Benjamin Van Durme. 2019. Improved lexically constrained decoding for translation and monolingual rewriting. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 839–850.

Shuyang Jiang, Yuhao Wang, and Yu Wang. 2023. Selfevolve: A code evolution framework via large language models. *CoRR*, abs/2306.02907.

Huayang Li, Guoping Huang, Deng Cai, and Lemao Liu. 2020. Neural machine translation with noisy lexical constraints. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1864–1874.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Sean Welleck, Bodhisattwa Prasad Majumder, Shashank Gupta, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. *CoRR*, abs/2303.17651.

Ian R McKenzie, Alexander Lyzhov, Michael Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan McLean, Aaron Kirtland, Alexis Ross, Alisa Liu, et al. 2023. Inverse scaling: When bigger isn't better. *arXiv preprint arXiv:2306.09479*.

Alexander Molchanov, Vladislav Kovalenko, and Fedor Bykov. 2021. Promt systems for wmt21 terminology translation task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 835–841.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models. In Mary Nurminen, Judith Brenner, Maarit Koponen, Sirkku Latomaa, Mikhail Mikhailov, Frederike Schierl, Tharindu Ranasinghe, Eva Vanmassenhove, Sergi Alvarez Vidal, Nora Aranberri, Mara Nunziatini, Carla Parra Escartín, Mikel L. Forcada, Maja Popovic, Carolina Scarton, and Helena Moniz, editors, *Proceedings of the 24th Annual Conference of the European Association for Machine Translation, EAMT 2023, Tampere, Finland, 12-15 June 2023*, pages 227–237. European Association for Machine Translation.

Yongyu Mu, Abudurexiti Reheman, Zhiquan Cao, Yuchun Fan, Bei Li, Yinqiao Li, Tong Xiao, Chunliang Zhang, and Jingbo Zhu. 2023. Augmenting large language model translators via translation memories. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10287–10299.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Silviu Pitis, Michael R. Zhang, Andrew Wang, and Jimmy Ba. 2023. Boosted prompt ensembles for large language models. *CoRR*, abs/2304.05970.

CCL 2024

Matt Post and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1314–1324.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2685–2702.

Dimitar Shterionov, Félix do Carmo, Joss Moorkens, Murhaf Hossari, Joachim Wagner, Eric Paquin, Dag Schmidtke, Declan Groves, and Andy Way. 2020. A roadmap to neural automatic post-editing: an empirical approach. *Machine Translation*, 34:67–96.

Kai Song, Yue Zhang, Heng Yu, Weihua Luo, Kun Wang, and Min Zhang. 2019. Code-switching for enhancing nmt with pre-specified translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 449–459.

Chau Tran, Shruti Bhosale, James Cross, Philipp Koehn, Sergey Edunov, and Angela Fan. 2021. Facebook ai wmt21 news translation task submission. In *Proceedings of the Sixth Conference on Machine Translation*, pages 205–215.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, pages 5998–6008.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. Context-aware monolingual repair for neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 877–886.

Thuy-Trang Vu and Reza Haffari. 2018. Automatic post-editing of machine translation: A neural programmer-interpreter approach. In *Empirical Methods in Natural Language Processing 2018*, pages 3048–3053. Association for Computational Linguistics (ACL).

Yuguang Wang, Shanbo Cheng, Liyang Jiang, Jiajun Yang, Wei Chen, Muze Li, Lin Shi, Yanfeng Wang, and Hongtao Yang. 2017. Sogou neural machine translation systems for wmt17. In *Proceedings of the Second Conference on Machine Translation*, pages 410–415.

Ke Wang, Shuqin Gu, Boxing Chen, Yu Zhao, Weihua Luo, and Yuqi Zhang. 2021a. Termmind: Alibaba's wmt21 machine translation using terminologies task submission. In *Proceedings of the Sixth Conference on Machine Translation*, pages 851–856.

Weixuan Wang, Wei Peng, Xupeng Meng, and Qun Liu. 2021b. Huawei aarc's submissions to the wmt21 biomedical translation task: Domain adaption from a practical perspective. In *Proceedings of the Sixth Conference on Machine Translation*, pages 868–873.

Shuo Wang, Peng Li, Zhixing Tan, Zhaopeng Tu, Maosong Sun, and Yang Liu. 2022a. A template-based method for constrained neural machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3665–3679.

Shuo Wang, Zhixing Tan, and Yang Liu. 2022b. Integrating vectorized lexical constraints for neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7063–7073.

Jinghui Yan, Jiajun Zhang, JinAn Xu, and Chengqing Zong. 2019. The impact of named entity translation for neural machine translation. In *Machine Translation: 14th China Workshop, CWMT 2018, Wuyishan, China, October 25-26, 2018, Proceedings 14*, pages 63–73. Springer.

Zixin Zeng, Rui Wang, Yichong Leng, Junliang Guo, Xu Tan, Tao Qin, and Tie-yan Liu. 2023. Extract and attend: Improving entity translation in neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1697–1710.

Jiacheng Zhang, Huanbo Luan, Maosong Sun, Feifei Zhai, Jingfang Xu, and Yang Liu. 2021. Neural machine translation with explicit phrase alignment. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1001–1010.

CCL 2024

| | CCR% | Cost(s) | CCR% | Cost(s) |
|---|---|---|---|---|
| *Direction* | *English-Chinese* | | *English-Russian* | |
| Iteration0 | 92.6 | 1.48 | 87.9 | 2.28 |
| Iteration1 | 93.9 | 2.21 | 93.0 | 3.26 |
| Iteration2 | 94.6 | 2.02 | 94.8 | 2.99 |
| Iteration3 | 95.0 | 2.39 | 95.1 | 3.17 |
| Iteration4 | 95.3 | 2.56 | 96.2 | 3.02 |
| Iteration5 | 95.6 | 2.42 | 96.9 | 3.25 |

Table 9: The processing time for each data point at various stages using the Qwen-14b-chat model on an A100 GPU.

| | CCR% | Cost($) | CCR% | Cost($) |
|---|---|---|---|---|
| *Direction* | *English-Chinese* | | *English-Russian* | |
| Iteration0 | 92.6 | 0.62 | 95.8 | 0.75 |
| Iteration1 | 95.1 | 1.20 | 97.0 | 1.28 |
| Iteration2 | 95.6 | 1.26 | 97.3 | 1.08 |
| Iteration3 | 95.9 | 1.31 | 97.5 | 1.23 |
| Iteration4 | 95.9 | 1.34 | 97.6 | 1.23 |
| Iteration5 | 95.9 | 1.34 | 97.6 | 1.21 |

Table 10: The monetary costs involved in processing every 1,000 data points using gpt3.5-turbo-0613 at different stages.

Chenrui Zhang, Lin Liu, Jinpeng Wang, Chuyuan Wang, Xiao Sun, Hongyu Wang, and Mingchen Cai. 2023a. PREFER: prompt ensemble learning via feedback-reflect-refine. *CoRR*, abs/2308.12033.

Huaao Zhang, Qiang Wang, Bo Qin, Zelin Shi, Haibo Wang, and Ming Chen. 2023b. Understanding and improving the robustness of terminology constraints in neural machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6029–6042.

Kechi Zhang, Zhuo Li, Jia Li, Ge Li, and Zhi Jin. 2023c. Self-edit: Fault-aware code editor for code generation. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 769–787. Association for Computational Linguistics.

## A Cost of Iterations

After undergoing revisions, our TAR significantly enhances the performance of translation systems in constrained translation scenarios. However, multiple rounds of iteration introduce additional computational and financial costs. To quantitatively assess these extra expenditures, we conducted evaluations on the WMT21 terminology translation dataset. Results of time cost are shown in Table 9, and the monetary costs are shown in Table 10. We repeated these tests 10 times for each stage and reported the average scores.

Our results indicate a trend of diminishing returns beyond the third iteration, both in terms of performance and cost-efficiency. Specifically, the time cost stabilizes at 10 seconds per data point and the monetary expense at approximately $4.5 per 1,000 data points by this iteration. Considering that TAR can significantly enhance the CCR in constrained translations , we believe that the cost is completely within an acceptable range.

## B Impact of the number of constraints

In conventional constrained translation methods, it has been observed that as the number of constraints within a single sentence increases, the CCR shows a decreasing trend. To investigate whether our method encounters the same challenge, we conducted experiments on the RTT (Zhang et al., 2023b) dataset. This dataset is composed of 500 samples meticulously selected by linguistic experts from the WMT 13-18 English-German translation test sets, with each sample being accompanied by at least six constraints. These constraints were chosen from a carefully curated set of noun phrases (e.g., names of organizations, individuals, movies, and brands) and common expressions.

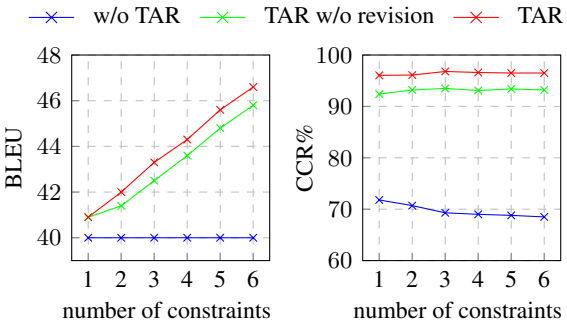



Figure 5: (a) The impact of increasing the number of constraints on BLEU. (b) The effect of increasing the number of constraints on CCR.

To simulate different numbers of constraints, we adopted a method similar to that described in (Zhang et al., 2023b), assuming each sentence in the test set corresponds to $N$ constraints, we randomly selected

| **Wikitionary** | **BLEU** | **COMET** | **CCR%** |
|---|---|---|---|
| LLM Const. Trans. | 32.1 | 87.3 | 97.2 |
| +Revision | 32.0 | 87.3 | 98.9 |
| **WMT21 TT en-zh** | **BLEU** | **COMET** | **CCR%** |
| LLM Const. Trans. | 36.4 | 86.9 | 92.6 |
| +Revision | 35.9 | 86.9 | 95.9 |
| **ETC en-ru** | **BLEU** | **COMET** | **CCR%** |
| LLM Const. Trans. | 32.6 | 89.9 | 88.8 |
| +Revision | 32.5 | 89.8 | 89.8 |

Table 11: Comparison of metrics before and after revision on the Wiktionary, WMT21 TT en-zh, and ETC en-ru datasets.

| **Constraints** | virus spread → 病毒传播 & Wuhan → 武汉 or 武汉市 |
|---|---|
| Source | In early and mid-January 2020, the virus spread to other Chinese provinces, helped by the Chinese New Year migration and Wuhan being a transport hub and major rail interchange. |
| Reference | 在2020年1月初至1月中旬，受中国春节人口大流动和武汉作为交通枢纽和主要铁路枢纽的影响，病毒传播到了中国其他省份。 |
| Const. Trans. | 2020年1月初和中旬，病毒通过中国春节迁徙和武汉作为交通枢纽和主要铁路换乘站的帮助，传播到其他中国省份。 |
| +Revision | 2020年1月初和中旬，病毒传播到其他中国省份，得益于中国春节迁徙和武汉市作为交通枢纽和主要铁路换乘站的地位。 |

Table 12: A case demonstrated that duringthe revsion phase,the constrain oftranslating “virus spread” to “病毒传播” was completed. Meanwhile, the model reorganized the sentence structure and adjusted the wording, resulting in a 5.25 decrease in BLEU score. However, the overall fluency did not change.

between 1 to $N$ constraints for testing. Consequently, we constructed $k$ test subsets, with the number of constraints ranging from 1 to $k$, Figure 5 presents the results for two metrics ($BLEU, CCR$) as the number of constraints ($k = 6$) varies. From the results, we can observe:

(1) As the number of constraints increases, the BLEU score of a standard LLM does not change, but the BLEU of both our proposed constraint-aware translator and TAR show a certain degree of improvement. When the number of constraints is 6, the BLEU of both methods can increase by about 5.5 points. This phenomenon is understandable because providing more constraints also means that more key parts of the translation sentence are already perceived by the model on how they should be translated.

(2) Similar to traditional constrained translation methods, using LLM as a translator alone, its CCR decreases as the number of constraints increases. However, the CCR of our proposed method does not change significantly, which also indicates that TAR can handle a greater number of constrained translation scenarios in real-world contexts.

## C Analysis of Causes for BLEU Score Decline During the Revision Stage

We observed that, among all twelve language directions, three exhibited a decline in BLEU scores after the revision stage. Theoretically, completing more constraints correctly during revision should lead to an increase in BLEU, a string-matching-based metric. However, in reality, to fulfill specific constraints, the model may employ different words or rephrase entire sentences to ensure semantic coherence, which can result in a decrease in BLEU scores. To investigate whether the revision stage could potentially degrade translation quality, we measured the changes in COMET scores before and after revision for these three language directions. The results, as shown in Table 11, indicate that although BLEU scores declined, the COMET scores remained stable, suggesting that TAR does not compromise overall translation quality. We provide a case study in Table 12 to further illustrate this phenomenon.

## D Applying mainstream constraint translation methods to LLMs

In the NMT era, many studies have explored training NMT models with data augmentation to develop constrained translation capabilities. Representative approaches include Code-switching, Append, etc. To investigate the effectiveness of these methods on LLMs, we performed experiments using the IATE and

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1059-1074, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

| **Method** | **BLEU** | **CCR%** | **BLEU** | **CCR%** |
|---|---|---|---|---|
| *Dataset* | *IATE* | | *Wiktionary* | |
| LLM Trans | 32.0 | 85.4 | 32.0 | 88.7 |
| LLM Code-switching | 31.8 | 94.0 | 31.9 | 92.6 |
| LLM Append | 31.6 | 93.3 | 31.5 | 91.9 |
| LLM Const. Trans. | 32.0 | 96.2 | 32.1 | 97.2 |
| +Revision | 32.0 | 98.9 | 32.0 | 98.9 |

Table 13: Performance of traditional data augmentation methods on LLMs.

Wiktionary datasets. The results are presented in Table 13. Although code-switching prompts are somewhat effective, they typically decrease translation quality and worsen CCR metrics compared to natural language prompts. We speculate that this is because LLMs have not encountered similar prompt formats during training, leading to alignment issues during task execution. Adopting techniques such as few-shot learning and fine-tuning may mitigate issues of misalignment between prompts and model training data. However, our proposed TAR method uses natural language prompts, which are more suitable for the processing style of large language models, thus enabling better understanding and execution of translation tasks.

CCL 2024

# A Multi-Task Biomedical Named Entity Recognition Method Based on Data Augmentation

**Hui Zhao[1], Di Zhao[1,2,3]*, Jiana Meng[1], Shuang Liu[1], Hongfei Lin[2]**
[1]School of Computer Science and Engineering, Dalian Minzu University, Liaoning, China
[2]School of Computer Science and Technology, Dalian University of Technology, Liaoning, China
[3]Postdoctoral workstation of Dalian Yongjia Electronic Technology Co., Ltd, Liaoning, China
zhaohui0817@163.com, zhaodi@dlnu.edu.cn, mengjn@dlnu.edu.cn
liushuang@dlnu.edu.cn, hflin@dlut.edu.cn

## Abstract

The rapid development of artificial intelligence has led to an explosion of literature in the biomedical field, and Biomedical Named Entity Recognition (BioNER) can quickly and accurately identify key information from unstructured text. This task has become an important topic to promote the rapid development of intelligence in the biomedical field. However, in the Named Entity Recognition (NER) of the biomedical field, there are always some problems of unclear boundary recognition, the underutilization of hierarchical information in sentences and the scarcity of training data resources. Based on this, this paper proposes a multi-task BioNER model based on data augmentation, using four data augmentation methods: Mention Replacement (MR), Label-wise token Replacement (LwTR), Shuffle Within Segments (SiS) and Synonym Replacement (SR) to increase the training data. The syntactic information is extracted by incorporating the input sentence into the Graph Convolutional Network (GCN), and then the tag information encoded by BERT is interacted through a co-attention mechanism to obtain an interaction matrix. Subsequently, NER is performed through boundary detection tasks and span classification tasks. Comparative experiments with other methods are conducted on the BC5CDR and JNLPBA datasets, as well as the CCKS2017 dataset. The experimental results demonstrate the effectiveness of the model proposed in this paper.

## 1 Introduction

Named Entity Recognition (NER) aims to identify named entities of specific meaning from the text and classify them into pre-defined entity types. In specific areas, such as biomedicine, NER needs to identify entity types, such as chemicals, diseases, genes, and proteins. NER plays a major role in critical information extraction and provides an important basis for other downstream tasks.

In deep learning methods, the forms of biomedical entities is more complex, usually composed of multiple words, and there are many abbreviations, whose forms are changeable and lack of clear rules to follow, which makes it difficult to determine the physical boundaries of entities and increases the difficulty of identification. Moreover, well-annotated datasets in the medical field are relatively scarce, further reducing the identification effect. Biomedical Named Entity Recognition (BioNER) also faces nested overlapping entities, leading to further increased difficulty in identification. Limited use of syntactic information in corporcorpus in BioNER methods, which also leads to the problem of poor results. Based on the above challenges, this paper presents a multi-task BioNER model based on data augmentation, with the following main contributions:

(1)The training data set is expanded by using Mention Replacement (MR), Label-wise token Replacement (LwTR), Shuffle Within Segments (SiS) and Synonym Replacement (SR), so as to solve the problem of scarce training data in the medical field and enhance the robustness of the model.

(2)The BioNER task is transformed from the traditional sequence annotation task to multi-task learning, including boundary detection and span classification joint task, thus solving the problems of unclear boundary recognition and entity nesting.

©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1075-1086, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

(3)The data set is used for syntactic analysis, and the Graph Convolutional Network (GCN) is used to encode the syntactic analysis graph and integrate it into the multi-task learning, so that the model can fully learn the syntactic components in the text and the relationship between the components, so as to improve the performance of NER.

## 2 Related Work

NER is one of the important indicators to evaluate the effect of information extraction. This paper introduces a data augmentation-based multi-task BioNER model. Next, data augmentation, syntactic analysis, and multi-task methods will be detailed to get a more comprehensive understanding of the study.

### 2.1 Data Augmentation Methods

Data augmentation is a method of increasing the amount of data based on different modification strategies, generating more and more diverse data sets by transforming, expanding or changing the original data. When training machine learning models, data augmentation acts as a regularization factor to avoid overfitting and increase the generalization of the model. Data augmentation can generate new training samples to simulate the diversity found in the real world, which helps the model learn more robust feature representations, thereby improving its performance on unseen data. Moreover, among the four data augmentation methods, the Shuffle within Segments method can generate adversarial examples that assist the model in better learning sentence features.

Data augmentation methods have been shown to be very effective in areas such as natural language processing. Wei et al.(Jason Wei and Kai Zou, 2019) enhanced the data by replacing the word with one of the synonyms retrieved from the English lexus. Kobayashi et al.(Sosuke Kobayashi, 2018) proposed a new method of labeled sentence data augmentation, called context augmentation. Wang et al.(Xinyi Wang et al., 2018) brought an extremely simple data augmentation strategy for neural machine translation (NMT): randomly replacing the words in the source and target sentences with other random words in the corresponding vocabulary, named this method as SwitchOut. Gao et al.(Fei Gao et al., 2019) proposed a novel neural machine translation data augmentation method to gently expands randomly selected words in sentences by mixing the context of multiple related words. Min et al.(Junghyun Min et al., 2020) explored several methods by applying syntactic transformations to sentences in the natural language inference (NLI) corpus to enhance the standard training set, increasing accuracy from 0.28 to 0.73. Xia et al.(Mengzhou Xia et al., 2019) used a bilingual dictionary and an unsupervised machine translation model to convert data from high resource languages to low resource languages to extend the machine translation training set of low resource languages.

### 2.2 Syntactic Analysis

In recent years, syntactic analysis is a task of Natural Language Processing (NLP), and has been successfully applied in many fields. In order to encode the context information of each word in a sentence, Kiperwasser(Kiperwasser E and Goldberg Y, 2016) introduces the Bidirectional Long Short-Term Memory (BiLSTM) into the syntactic analysis system based on transfer model and graph model, respectively. Wei et al.(Wei Wei et al., 2024) proposed a text similarity calculation method based on syntactic analysis, which summarizes the principles of other methods. Ding et al.(Ding Meirong et al., 2024) proposed to build a syntactic dependency tree of text first, introduce richer feature information to the model, and use the feature aggregation module to process the feature aggregation, which improved the accuracy of the model's prediction of word emotion analysis.

### 2.3 Multi-Task Learning Method

Multi-Task Learning (MTL) is a machine learning paradigm that aims to improve the performance of models by learning multiple related tasks simultaneously. Existing studies have begun to explore the association of NER tasks with other tasks and propose methods based on multi-task learning. Collobert et al.(Collobert R et al., 2011) trained a Window / Sentence method network to jointly perform NER, POS,

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1075-1086, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Chunk, and SRL tasks. Rei(Rei M, 2017) found that the performance of sequence annotation model can be improved by adding unsupervised language modeling objectives to the training process. Zhang(Zhang Xue et al., 2024) et al. proposed a multi-task classification algorithm based on loss optimization of gradient amplitude direction adjustment to transform multiple classification tasks into multiple dichotomy tasks. Song et al.(Song Donghuan et al., 2024) proposed to construct a novel classification model, which demonstrated the effectiveness of the proposed model by designing multi-task experiments and transfer learning experiments. Wang et al.(Wang Jingwei et al., 2024) proposed a multi-task Transformer model for flight state parameter regression and flight state classification, and the experimental results proved the superiority of the proposed model.

## 3 Method

First, the model uses four methods for data augmentation, with input data and label information processed through BERT encoding, and syntactic information extracted using GCN. New input data and label information, along with syntactic information via an attention interaction matrix, are then applied to the boundary detection and span classification tasks for entity recognition. The BioNER task is transformed from a sequence annotation task into a multi-task framework, specifically addressing the joint tasks of boundary detection and span classification, which solves the problems of unclear boundary recognition and entity nesting. Syntactic analysis is performed on the dataset, and GCN is used to encode the syntactic analysis graph, integrating it into the multi-task framework. This allows the model to fully learn the syntactic components in the text and their interrelationships, thereby improving the performance of NER. The overall architecture of the model is shown in Figure 1.

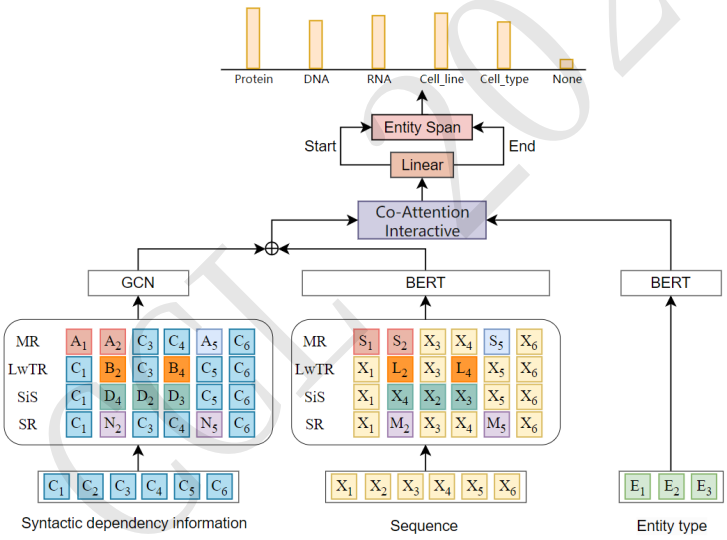



Figure 1: The overall architecture of the model

### 3.1 Data Augmentation Methods

In this paper, four methods of MR, LwTR, SiS and SR are used to change the input data, and these four methods are used for data augmentatio.

**Mention Replacement (MR)**:For each entity in the input sentence, a binomial distribution is used to randomly decide whether it should be replaced. The binomial distribution represents the resulting probability distribution of a series of independent repeated binary experiments. For example, in Figure 2, the chemical type entity ”22-oxacalcitriol [B-chem I-chem I-chem]” was replaced by ”Puromycin aminonucleoside [B-chem I-chem]”, an entity of the disease type ”secondary hyperparathyroidism [B-dise I-disorder]” was replaced by ”Cauda equina syndrome [B-dise I-dise I-disorder]”, ”low bone turnover [B-dise I-dise I-disease]” was replaced by ”neurotoxicity [B-disease]”, and ”renal failure [B-dise I-disorder]” was replaced by ”neurological deficits [B-dise I-disorder]”. The new training data, after entity

replacement, along with the original training data, are encoded by the pre-trained model BERT, and then decoded by the CRF to obtain the final label.

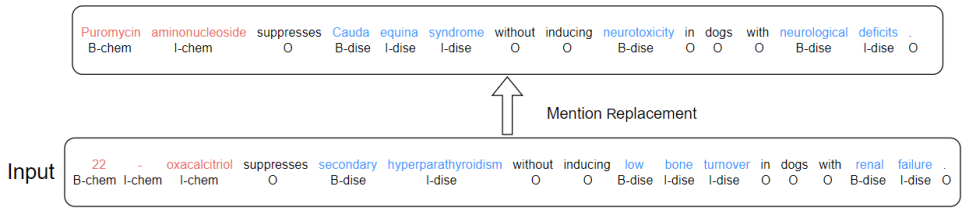



Figure 2: The Mention Replacement method

**Label-wise Token Replacement (LwTR)**: For each word, a binomial distribution is also used to randomly decide whether it should be replaced. If the probability result is replacement, use the tag-word distribution (the label-token) distribution built from the original training set to randomly select another word with the same tag and ultimately keep the original tag sequence unchanged. As shown in Figure 3, six words (suppresses, without, inducing, bone, dogs, and renal) were replaced by other words (aminonucleoside, levels, of, none, haven, hypercalcemia) with the same label as the original word.

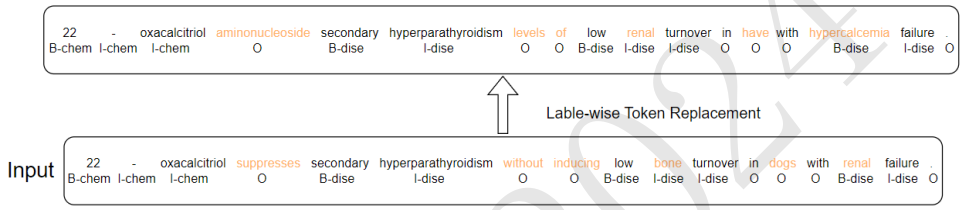



Figure 3: The Label-wise Token Replacement method

**Shuffle Within Segments (SiS)**: First, the word sequence is split into small parts of the same label, where each part corresponds to a word or a series of words of the same type. As shown in Figure 4,secondary hyperparathyroidism, without causing and low bone transformation all three phrases were disrupted, (secondary hyperparathyroidism) were disrupted as (hyperparathyroidism secondary), (without inducing) as (inducing without), and (low bone turnover) as (turnover low bone).

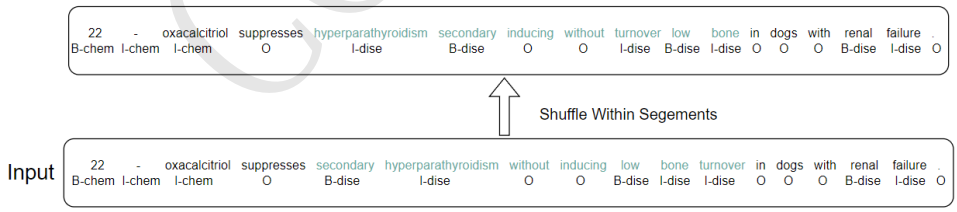



Figure 4: The Shuffle within Segments method

**Synonym Replacement (SR)**: The SR method is similar to LwTR, replacing words with synonyms retrieved using WordNet or Chinese WordNet. But the retrieved synonyms may include more than one word, so for BIO labels, a simple rule can be derived: If the replaced word is the first word in the entity (i.e., the corresponding label is "B-chem"), we assigned the same label to the first marker of the retrieved multi-word synonyms and assigned "I-chem" to the other words. If the replaced word is in an entity (i.e., the corresponding label is "I-chem"), we assign its label to all words of multi-word synonyms. As shown in Figure 5, (suppresses) replaced with (restrain), (low bone turnover) replaced with (suppression of bone turnover), and (renal failure) replaced with (impaired renal failure).

CCL 2024

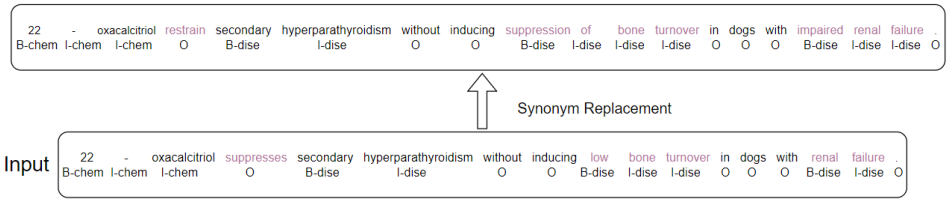



Figure 5: The Synonym Replacement method

The four methods shown above are used to enhance the input training data. These four methods use the binomial distribution to randomly replace and change. For the binomial distribution, the probability of success in each experiment is $p$, and the probability of failure is $q = 1 - p$, the probability quality function of the binomial distribution of $n$ experiments is shown as eq 1:

$$P(x = k) = C(n, k) \times p^k \times (1 - p)^{n-k} \tag{1}$$

$P(x = k)$ is the probability of a successful event $k$, $C(n, k)$ is the number of combinations, indicating the combination of $k$ times in $n$ experiments, $p$ is the probability of success in each experiment, $q(q = 1 - p)$ is the probability of failure in each experiment, $n$ is the total number of independent experiments, $k$ is the number of successful events. A binomial distribution was used each time to calculate whether substitutions were needed.

### 3.2 Encoding Layer

**Syntactic Dependency Tree**: Syntactic Dependency Tree is a type of syntactic structure tree that illustrates the grammatical relationships between each word in a sentence and how they are connected within the sentence. There is typically a root node that represents the core of the entire sentence, with other words connected to the root node or to each other through various types of dependency relationships.

**GCN**: GCN is a deep learning model used to process graph data. In NLP, text data is modeled into a graph where each word is represented by a node in the graph, and the relationships between words, such as dependencies or co-occurrences, are represented as edges in the graph. The context of a named entity typically contains key information, and the GCN can model each node by considering the neighboring nodes and can capture the dependencies between words. In this paper, the GCN is used to extract syntactic information from the input sentences.

The GCN makes the input information and its dependent syntactic tree interact, so that the model fully learns the existing dependent relationship in the sentence. The $h1$ is the input information encoded by BERT, $A$ is the adjacency matrix, $D$ is the degree matrix, and $w$ and $b$ are the learnable parameters. The GCN extracts the syntactic features in the sentence as shown in formula eq 2:

$$G_x = RELU(DAh_1w + b) \tag{2}$$

For the training data augmented with the four data augmentation methods, BERT is used for encoding. In addition to the syntactic information in the sentences, entity-specific information is also encoded within the entity types. To better represent the entity type label information, average pooling is conducted on the BERT-encoded entity type labels, resulting in rich representations of these labels.

### 3.3 Fusion Layer

The co-attention mechanism is an attention method that processes sequence data and establishes inter-sequence associations. It calculates the weighted attention for each element, making the information transmission of long sequences more efficient. This mechanism enables the model to consider all elements in the input sequence for both encoding and decoding, thereby enabling the model to better understand the long-distance dependencies between sequences. In this paper, we use the co-attention mechanism to learn syntactic information in sentences for improved feature fusion.

CCL 2024

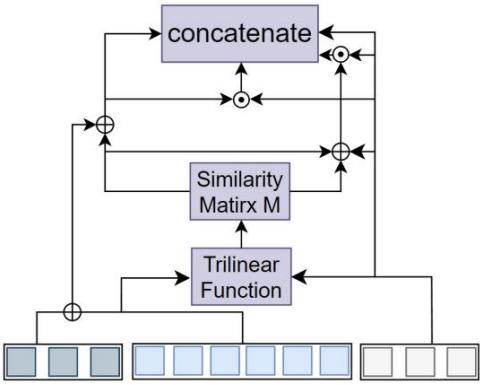



Figure 6: Co-Attention Mechanism Model

In order to better perform feature fusion, the common attention interaction mechanism is used to fuse features. As shown in Figure 6, in the output of the GCN and label after mean pooling output, using the common attention interaction network, make the syntactic information contained in the input sentences and sentences and the entity label contains specific information fully interaction, makes the model can better learn the potential interaction information and specific information, which can better represent the entity span boundary, boundary information as shown in eq 3:

$$M = w^T[h_x, G_x, h_e] \tag{3}$$

Where $w$ is the trainable weight, $M$ is the final attention weight matrix. By performing row-wise and column-wise operations, we respectively obtain the interaction matrices $Mr$ for entity boundaries and $Mc$ for entities. By using eqs 4 and 5, we obtain the interaction between labels and tokens, and tokens and labels.

$$H_l^t = M_r \bullet h_e \tag{4}$$

$$H_t^l = M_r \bullet M_c^T \bullet h_x \tag{5}$$

### 3.4 Classification Layer

The multi-task NER model allows for sharing and utilizing data across multiple tasks. The method adopted in this paper, which employs boundary detection and span classification, is capable of identifying nested entities. The boundary detection component ascertains the specific boundaries of nested entities. Multi-task learning reduces the complexity of maintaining and debugging individual models, leading to decreased computational resource requirements and improved system efficiency. The specific method is shown in Figure 7.

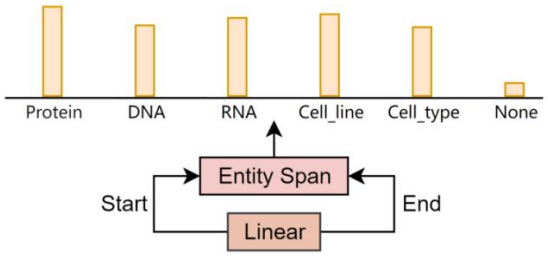



Figure 7: Multi-Task Learning Model

The head and tail positions of entities are identified first, and then the label classification task of entity type is performed. Head and tail entity recognition tasks require the model to perform more fine-grained identification near the boundaries of entities, thus providing more contextual information and being more helpful to accurately determine entity boundaries. The entity classification task identifies entity types,

CCL 2024

which provides additional information about the entity content. This multitasking method allows the model to understand entities from different perspectives and improves the accuracy of entity identification; the model can better determine the location of the head and tail entities, and the information of the head and tail entities can help each other to improve performance and overall performance.

Specifically, calculating the start or end position of the entity span, as the front task of multi-task mode, calculating the correct entity boundaries, laying the foundation for the following entity type classification task. During training, the cross-entropy loss was used to optimize the boundary of the entity span. The formula is as eqs(6;7;8):

$$B = [h_x; H_l^t h_x \bullet H_l^t; h_x \bullet H_t^l] \tag{6}$$

$$p(\theta)_i = \frac{exp(w_1^T B_i)}{\sum_j exp(w_1^T B_j)} \tag{7}$$

$$L_{ht}(\theta) = -\sum_{i=1}^{N} \hat{y}_i log P(\theta)_i \tag{8}$$

The span classification task involves matching the head and tail labels of each boundary tag to classify entity spans into corresponding semantic labels. If a candidate entity span is not an entity, it is labeled as the "None" category. An entity span $V_{span}$ is defined as eq 9:

$$V_{span} = [B_i + B_j; B_i - B_j] \tag{9}$$

Finally, the entity span representation is sent to a softmax layer to predict the probability $P_{span}$ of entity labels. The entity label prediction loss $L_{span}$ is minimized through eqs 10 and 11.

$$P_{span}(\theta) = softmax(MLP(w^T V_{span} + b)) \tag{10}$$

$$L_{span}(\theta) = -\sum_{i=1}^{c} (y_{span}^{\Delta i}) log(P_{span}^{i}(\theta)) \tag{11}$$

In the multi-task approach, during the training phase, the loss of entity type labels and boundaries is minimized to the maximum extent possible using eq 12.

$$L(\theta) = L_{ht}(\theta) + L_{span}(\theta) \tag{12}$$

## 4 Experiments and Analysis

### 4.1 Experimental Data

The statistics of the datasets are shown in Table 1:

| | JNLPBA | BC5CDR | CCKS2017 |
|---|---|---|---|
| Train | 14,690 | 15,935 | 1,561 |
| Dev | 3,856 | 4,012 | 334 |
| Test | 3,856 | 4,012 | 334 |

Table 1: Dataset Statistics

### 4.2 Experimental Results and Analysis

**Comparative Experiments of Data Augmentation**: In order to verify the effectiveness of data augmentation methods, this study conducted experiments on the BC5CDR dataset, the JNLPBA dataset and the CCKS2017 dataset using single data augmentation methods as well as the combination of all four methods under the sequence labeling method. Comparative experiments were also conducted with other methods, and the experimental results are presented in Table 2 :

CCL 2024

| Model | BC5CDR | | | JNLPBA | | | CCKS2017 | | |
|---|---|---|---|---|---|---|---|---|---|
| | P(%) | R(%) | F1(%) | P(%) | R(%) | F1(%) | P(%) | R(%) | F1(%) |
| BERT+CRF | 87.0 | 81.5 | 84.2 | 71.8 | 77.3 | 74.5 | 88.7 | 93.0 | 90.9 |
| BERT+BiLSTM+CRF | 87.7 | 82.0 | 84.8 | 71.0 | 79.6 | 75.0 | 90.9 | 91.6 | 91.2 |
| BERT+CNN+CRF | 88.7 | 82.6 | 85.5 | 71.2 | 80.2 | 75.4 | 91.3 | 92.2 | 91.7 |
| BERT+CRF+MR+GCN | 89.2 | 87.8 | 88.5 | 74.6 | 79.8 | 77.2 | 92.2 | 92.8 | 92.5 |
| BERT+CRF+LwTR+GCN | 88.6 | **89.2** | 88.9 | 74.7 | 80.3 | 77.5 | 92.5 | 92.9 | 92.7 |
| BERT+CRF+SiS+GCN | 89.5 | 88.9 | 89.1 | 77.8 | 78.0 | 77.9 | 92.7 | 93.7 | 93.2 |
| BERT+CRF+SR+GCN | 89.1 | 88.8 | 89.0 | 74.2 | **81.0** | 77.6 | 92.4 | 93.4 | 92.9 |
| BERT+CRF+ALL+GCN | **90.2** | 88.7 | **89.4** | **78.5** | 77.9 | **78.2** | **93.2** | **93.8** | **93.5** |

Table 2: Baseline Comparison Experiments Results

According to the experimental results shown in Table 2, the model with the added syntactic information exhibits a significant improvement in F1 score, whether using single data augmentation method or all four data augmentation methods combined. Among the four methods of MR, LwTR, SiS and SR, it can be observed that in the experimental results of the three datasets, SiS method performs the best, followed by SR, then LwTR, and finally MR. The reason why the SiS method performs the best is because it increases diversity. Shuffling disrupts the original order of the text, enhancing the diversity of the dataset, which helps the model to learn the associations between words and phrases in different contexts more comprehensively, thus improving its generalization ability and robustness. The advantage of the SR method over other methods lies in its ability to alleviate the problem of data sparsity. Some words may appear less frequently in the dataset, and synonym replacement can increase the number of samples for these words. The relatively poor performance of the LwTR method and the MR method is due to the fact that the replaced words introduce new meanings for the model to learn. However, for the model, the relationships between each word in the context and the sentence are relatively more important. After incorporating syntactic information and data augmentation, the F1 scores of the BC5CDR dataset and the JNLPBA dataset improved by 3.9% and 2.8% respectively compared to the baseline experimental method. On the CCKS2017 dataset, the F1 score improved by 1.8%. After incorporating data augmentation and syntactic information, the model achieved its best performance. The experimental results demonstrate that combining various data augmentation techniques and syntactic information can significantly improve the performance of NER models.

**Comparative Experiments of Multi-Task Learning**: Experiments were conducted on the BC5CDR dataset, the JNLPBA dataset and the CCKS2017 dataset using the multi-task method, and comparative experiments were conducted with other methods. The experimental results are shown in Tables 3 and 4:

| Model | BC5CDR | | | Model | JNLPBA | | |
|---|---|---|---|---|---|---|---|
| | P(%) | R(%) | F1(%) | | P(%) | R(%) | F1(%) |
| CASN(2021) | 89.5 | 90.9 | 90.2 | CASN(2021) | 75.8 | 80.0 | 78.0 |
| BERTForTC (2022) | - | - | 90.8 | SciFive(2021) | - | - | 77.5 |
| UniNER (2023) | - | - | 89.3 | MINER (2022) | - | - | 77.0 |
| GoLLIE (2023) | - | - | 88.4 | BioDistilBERT(2023) | 73.5 | **85.5** | 79.1 |
| Ours | **92.7** | **91.5** | **92.1** | Ours | **78.9** | 82.1 | **80.5** |

Table 3: Comparative Experimental Results on BC5CDR and JNLPBA

| Model | CCKS2017 | | |
|---|---|---|---|
| | P(%) | R(%) | F1(%) |
| Yu et al. (2022) | - | - | 93.8 |
| Lin et al. (2023) | 92.6 | 91.9 | 92.2 |
| Li et al. (2023) | 87.4 | 92.1 | 89.7 |
| Shen et al. (2023) | 93.3 | 92.5 | 92.6 |
| Ours | **94.1** | **96.3** | **95.2** |

Table 4: Comparative Experimental Results on CCKS2017

As shown in Tables 3 and 4, after comparative experiments with other methods, the proposed multi-task learning model for BioNER based on data augmentation outperforms other methods. It achieved the best F1 scores on both the BC5CDR dataset and the JNLPBA dataset, with improvements of 1.3% and 1.4% respectively. Similarly, it also achieved the best F1 score on the CCKS2017 dataset, with an improvement of 1.4%. This demonstrates the effectiveness of data augmentation, incorporation of syntactic information, and feature extraction using GCN. The model has successfully learned more dependencies existing in sentences, and its performance in boundary detection and span classification multi-task learning is excellent.

**Ablation Experiments**: In order to verify the effectiveness of the multi-task module, syntactic information, and data augmentation methods on the model's performance, we conducted ablation experiments on the BC5CDR dataset, the JNLPBA dataset and the CCKS2017 dataset. The experimental results are shown in Table 5:

| Model | BC5CDR | | | JNLPBA | | | CCKS2017 | | |
|---|---|---|---|---|---|---|---|---|---|
| | P(%) | R(%) | F1(%) | P(%) | R(%) | F1(%) | P(%) | R(%) | F1(%) |
| Ours | **92.7** | 91.5 | **92.1** | **78.9** | **82.1** | **80.5** | **94.1** | **96.3** | **95.2** |
| W/O MT | 90.2 | 88.7 | 89.4 | 78.5 | 77.9 | 78.2 | 93.2 | 93.8 | 93.5 |
| W/O GCN | 91.3 | 90.1 | 90.7 | 78.4 | 79.4 | 78.9 | 93.5 | 94.3 | 93.9 |
| W/O DA | 90.2 | **92.4** | 91.3 | 78.0 | 81.7 | 79.8 | 93.9 | 95.6 | 94.7 |

Table 5: Comparative Results of Ablation Experiments

As shown in Table 5, it is evident that removing either the multi-task module, syntactic information, or data augmentation leads to a decrease in the model's performance. After removing the multi-task module, syntactic information, and data augmentation, the F1 scores for the BC5CDR dataset decreased by 2.7%, 1.4%, and 0.8% respectively. For the JNLPBA dataset, the F1 scores decreased by 2.3%, 1.6%, and 0.7% respectively. Similarly, for the CCKS2017 dataset, the F1 scores decreased by 1.7%, 1.3%, and 0.5% respectively. The experimental results demonstrate that the multi-task module has the greatest impact on the model, followed by syntactic information, and finally data augmentation methods. Using the multi-task learning approach with entity boundary detection and span classification tasks enables the recognition of all entities and effectively addresses the out-of-vocabulary problem. This is crucial for NER tasks as it ensures comprehensive entity recognition. The syntactic relationships provided by syntactic information serve as shallow features and have a beneficial effect on NER. Data augmentation methods enhance the robustness of the model, albeit to a lesser extent compared to other factors. The experimental results provide ample evidence of the positive impact and effectiveness of the multi-task learning approach, syntactic information, and data augmentation methods on the model.

CCL 2024

### 4.3 Analysis of Errors

On the BC5CDR dataset, the CASN method is a BioNER model based on a multi-task approach. BioLinkBERT utilizes inter-document connections for self-supervised training. BERTForTC is a BERT-based model specifically designed for text classification tasks. UniNER leverages the large model GPT to enhance the performance of the NER model. GoLLIE utilizes a large model guide for information extraction to improve model effectiveness, showing significant results in NER.

On the JNLPBA dataset, SciFive is a domain-specific T5 model that has been pre-trained on extensive biomedical corpora and has delivered superior performance in the context of NER models. MINER proposes a new learning framework for NER, approaching the challenge from an information-theoretic viewpoint and addressing the issue of out-of-vocabulary (OOV) entities. BINDER improves the efficiency of NER by applying contrastive learning to project candidate text spans and entity types into a shared vector representation space. The BioDistilBERT method, based on the DistilBERT model, has been fine-tuned on biomedical text and has yielded excellent outcomes.

On the CCKS2017 dataset, Yu et al. proposed a multi-task learning approach for entity recognition, which recognizes entities through the interaction of two tasks: entity recognition and participle training. The two tasks are not both NER in this method, while the boundary recognition and entity type classification proposed in this paper are both related to NER, and the experimental results also proved that the proposed method is better. Duan et al. tested the micro-F1 values and the changes of the F1 values of each entity after augmenting the word embedding with Chinese character features, and experimented the effect of different methods on the CCKS2017 dataset. Lin et al. proposed a pre-trained model multi-layer dynamic fusion method. The multi-layer dynamic fusion method can fully utilize the knowledge of each layer of the pre-trained model, so that the results contain rich syntactic, semantic and other feature information, improve the representation ability of the model in the task, and enhance the flexibility of the model. Li et al. proposed a novel entity recognition method combined with CNN. The method introduces an external vocabulary through SoftLexicon and combines character and vocabulary vectors to enhance character feature representation using word frequency normalization of associated characters. The experimental results show that this method outperforms SoftLexicon on multiple datasets, effectively improving the performance of NER. Shen et al. proposed a model that combines whole-word masking and BERT pre-processing, extracting features through BiLSTM and attention mechanisms, and using CRF decoding to enhance the recognition effect of Chinese medical text. The experimental results on different datasets show that the model achieved high F1 scores in the task of BioNER, proving its effectiveness.

These methods have overlooked the shallow-level features inherent in sentences and also did not have an adequate amount of training data for thorough training. In this paper, the model has learned the shallow-level features within sentences and has been enhanced with data augmentation methods to strengthen the training data. The results have also confirmed the effectiveness of this model.

## 5 Summary and Outlook

This paper proposes a multi-task BioNER model based on data augmentation. It employs four data augmentation methods, MR, LwTR, SiS and SR, to augment the training data. The model utilizes GCN to extract syntactic features from sentences. These multiple features are then interacted through a co-attention mechanism and combined with boundary detection and span classification in a multi-task mode for NER. The experimental results confirm the effectiveness and positive impact of adopting multi-task learning and data augmentation methods.

In the future, we will explore transfer learning and large language models. Large language models, with more parameters and deeper structures, have the capability to capture richer and more complex data patterns and features. This enables them to learn more abstract and advanced representations, leading to better performance across various tasks.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1075-1086, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

## Acknowledgements

This work was supported by the Fundamental Research Funds for the Central Universities (0854-53) and the Natural Science Foundation of Liaoning Province, China (2022-BS-104).

## References

Chen, P., Peng, et al.: *Co-Attentive Span Network with Multi-Task Learning for Biomedical Named Entity Recognition.* In: 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2021, pp. 649-652.

COLLOBERT R,WESTON J,BOTTOU L,et al. *Natural language processing(almost) from scratch* [J]. Journal of Machine Learning Research, 2011,12:2493-2537.

Ding Meirong, Lai Jin Qian, Zeng Biqing, etc. *Level-level sentiment analysis based on local global context guidance* [J]. Software Guide, 2024,23 (01): 190-196.

Fei Gao, Jinhua Zhu, Lijun Wu, Yingce Xia, Tao Qin, Xueqi Cheng, Wengang Zhou, and Tie-Yan Liu. 2019. *Soft contextual data augmentation for neural machine translation.* In ACL, pages 5539–5544, Florence, Italy.

Jason Wei and Kai Zou. 2019. *EDA: Easy data augmentation techniques for boosting performance on text classi- fication tasks.* In EMNLP-IJCNLP, pages 6382–6388, Hong Kong, China.

Junghyun Min, R. Thomas McCoy, Dipanjan Das, Emily Pitler, and Tal Linzen. 2020. *Syntactic data augmentation increases robustness to inference heuristics.* In ACL, pages 2339–2352, Online.

KIPERWASSER E,GOLDBERG Y. Simple and Accurate Dependency Parsing Using Bidirectional LSTM Feature Representations [J].TACL,2016,4:313–327.

Kocaman, V., Talby, D.: *Accurate Clinical and Biomedical Named Entity Recognition at Scale.* Software Impacts 13, 100373 (2022).

Li Mingjian, Li Weijun, Wang Hairong. *Identification of fused association information with CNN* [J]. Journal of Zhengzhou University (Science edition), 2023,55 (05): 53-59.

Lin Lingde, Liu Na, Xu Zhenshun, et al. *Chinese medical named entity identification based on multi-layer dynamic fusion* [J]. Computer Engineering and Application, 2023:1-13.

Mengzhou Xia, Xiang Kong, Antonios Anastasopoulos, and Graham Neubig. 2019. *Generalized data augmentation for low-resource translation.* In ACL, pages 5786–5796, Florence, Italy.

Phan L N, Anibal J T, Tran H, et al. *SciFive: a text-to-text transformer model for biomedical literature* [J]. 2021.

REI M. *Semi-supervised multitask learning for sequence labeling* [C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017:2121-2130.

Rohanian, O., Nouriborji, M., Kouchaki, S., et al.: *On the Effectiveness of CompactBiomedical Transformers.* Bioinformatics 39 (3), btad103 (2023).

Sainz O, García-Ferrero I, Agerri R, et al. *GoLLIE: Annotation Guidelines improve Zero-Shot Information-Extraction* [C]//The Twelfth International Conference on Learning Representations. 2023.

Shen Tongping, Yu Lei, Shu Jianhua, etc. *Medical text entities based on the multi-head attention mechanism* [J]. Journal of Anqing Normal University (Natural Science Edition), 2023,29 (01): 85-91.

Song Donghuan, Hu Maodi, Ding Jielan, etc. *Research on cross-type text classification techniques based on multi-task learning* [J/OL]. Data analysis and knowledge discovery: 1-19 [2024-03-20].

Sosuke Kobayashi. 2018. *Contextual augmentation: Data augmentation by words with paradigmatic relations.* In NAACL, pages 452–457, New Orleans, Louisiana.

Wang Jingwei, Gao Yan, Song Xing, etc. *Aircraft state prediction based on the Transformer* [J]. Computer Engineering and Design, 2024,45 (02): 477-483.

Wang, X., Dou, S., Xiong, L., Zou, Y., Zhang, Q., Gui, T., Qiao, L., Cheng, Z., Huang, X.: *MINER: Improving Out-of-Vocabulary Named Entity Recognition from an Information Theoretic Perspective.* In:

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022, vol. 1, pp. 5590-5600.

Wei Wei, lilac incense, Guo Mengxing, etc. *Summary of text similarity calculation methods [J/OL].* Computer Engineering: 1-19 [2024-03-20].

Xinyi Wang, Hieu Pham, Zihang Dai, and Graham Neubig. 2018. *SwitchOut: an efficient data augmentation algorithm for neural machine translation.* In EMNLP, pages 856–861, Brussels, Belgium.

Yu, P., Chen, Y., Xu, J., et al.: *Entity Recognition Method for Electronic Medical Records Based on Multi-task Learning.* In: Computer and Modernization, 2022 (09), 40-50.

Zhang Xue, Tian LAN, Zeng Ming, etc. *An ECG signal multitask classification algorithm based on gradient amplitude direction adjustment* [J/OL]. Computer Science: 1-10 [2024-03-20].

Zhou W, Zhang S, Gu Y, et al. *UniversalNER: Targeted Distillation from Large Language Models for Open Named Entity Recognition* [C]//The Twelfth International Conference on Learning Representations. 2023.

CCL 2024

# Biomedical Event Causal Relation Extraction by Reasoning Optimal Entity Relation Path

Lishuang Li[1,*], Liteng Mi[1], Beibei Zhang[2], Yi Xiang[1], Yubo Feng[1],
Xueyang Qin[1], Jingyao Tang[1]
[1]School of Computer Science and Technology, Dalian University of Technology,
Dalian, Liaoning, China
[2]School of Computer Science and Technology,
Shandong Institute of Petroleum and Chemical Technology,
Dongying, Shandong, China
lils@dlut.edu.cn,{militn,lulu163163,xiangyi_job}@163.com,
argamx@126.com,qinxueyang@snnu.edu.cn,
tangjingyao@mail.dlut.edu.cn

## Abstract

Biomedical Event Causal Relation Extraction (BECRE) is an important task in biomedical information extraction. Existing methods usually use pre-trained language models to learn semantic representations and then predict the event causal relation. However, these methods struggle to capture sufficient cues in biomedical texts for predicting causal relations. In this paper, we propose a Path Reasoning-based Relation-aware Network (PRRN) to explore deeper cues for causal relations using reinforcement learning. Specifically, our model reasons the relation paths between entity arguments of two events, namely entity relation path, which connects the two biomedical events through the multi-hop interactions between entities to provide richer cues for predicting event causal relations. In PRRN, we design a path reasoning module based on reinforcement learning and propose a novel reward function to encourage the model to focus on the length and contextual relevance of entity relation paths. The experimental results on two datasets suggest that PRRN brings considerable improvements over the state-of-the-art models.

## 1 Introduction

Biomedical Event Causal Relation Extraction (BECRE) is an important task in the field of biomedical information extraction which aims to identify the causal relation between two events in biomedical texts. As shown in Fig.1, given the context and two events, a BECRE model needs to predict whether there is a causal relation between head and tail events, i.e. "*phosphorylated* $\xrightarrow{cause}$ *degraded*". BECRE has played an important role in many downstream tasks such as knowledge question answering (Mutabazi et al., 2021), automatic decision-making, and knowledge discovery (Cañizares-Díaz et al., 2021).

BECRE is challenging because event causal relations in biomedical texts are always expressed in implicit ways. Existing methods usually utilize pre-trained language models to predict event causal relations (Akkasi and Moens, 2021). Liang et al. (2022) analyzed the impact of in-domain pre-training and distillation on the performance of BERT (Kenton and Toutanova, 2019). Akkasi and Moens (2021) used four different biomedical repositories that are publicly available to investigate the effect of the corpus used for pre-training BioBERT (Lee et al., 2020). These methods solely learned features of context relying on pre-trained language models. For further enhancing the ability of models to understand the context, Zhang et al. (2023) introduced external knowledge to build hierarchical knowledge graphs

*Corresponding author
©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1087-1098, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

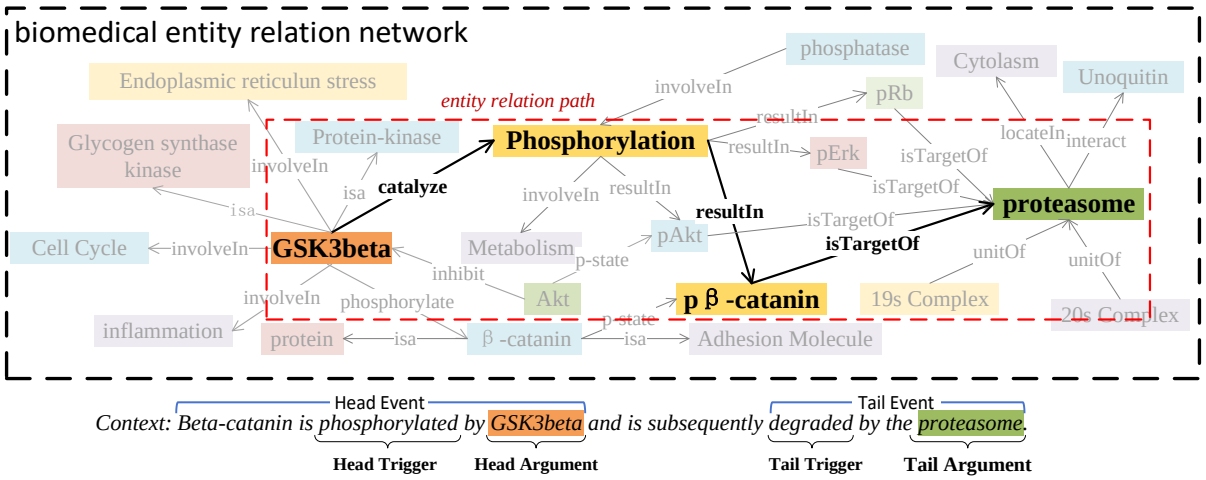



Figure 1: An illustration of expression of causal relations in biomedical texts and the entity relation path.

for the contexts. However, the above methods still have difficulty in capturing sufficient cues for event causal relations, leading to limited performance in BECRE.

To address the above issues, we pay attention to the entity relation path between biomedical events, and propose a method based on the entity relation path, named Path Reasoning-based Relation-Aware Network (PRRN). In PRRN, the entity relation path provides rich cues for predicting event causal relations. For example, in Fig.1, *"GSK3beta* $\xrightarrow{catalyze}$ *Phosphorylation* $\xrightarrow{resultIn}$ *pβ − catanin* $\xrightarrow{isTargetOf}$ *proteasome"* is a entity relation path between the event pair *(phosphorylated, degraded)*. The entity relation path describes multi-hop interactions between the head argument, *GSK3beta* and the tail argument, *proteasome*, including that *GSK3beta* can catalyze *Phosphorylation*, and *pβ-catanin* is the product of *Phosphorylation* and the target of *proteasome*. The interactions reveal an underlying cue that *β-catanin* is degraded by *proteasome* because *GSK3beta* converts *β-catanin* to *pβ-catanin* through *Phosphorylation*. So we can see that the entity relation path introduces multi-hop interactions of entities to provide rich cues for predicting event causal relation.

To obtain the optimal entity relation path, we design a path reasoning module based on reinforcement learning and propose a novel reward function. In the module, a policy-based agent network gradually approaches the optimal reasoning strategy in the way of Monte-Carlo Policy Gradient (Williams, 1992). The reward function we proposed in PRRN jointly encourages the length and contextual relevance of paths.

We conduct extensive experiments to evaluate the effectiveness of PRRN. The experimental results show that our method outperforms previous methods. In summary, the main contributions of this paper are:

- We propose a Path Reasoning-based Relation-Aware Network (PRRN) for BECRE, which explores richer cues to predict event causal relations by reasoning relation paths between entity arguments of biomedical events, i.e. entity relation paths.

- We design a path reasoning module based on reinforcement learning to reason optimal entity relation paths and propose a novel reward function to encourage the model to focus on the length and contextual relevance of entity relation paths.

- We conduct a series of experiments to verify the effectiveness of PRRN. The experimental results show that the entity relation path reasoned using reinforcement learning are beneficial for predicting event causal relation. Moreover, our method outperforms previous state-of-the-art methods on two datasets.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1087-1098, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

## 2 Related Work

Early studies of BECRE primarily relied on manually defined rules or shallow neural networks for event causal relation prediction (Mihăilă and Ananiadou, 2014; Hahn-Powell et al., 2016). Hahn-Powell et al. (2016) conducted rule-based and LSTM-based methods to practice BECRE. However, rule-based methods attempt to induce extremely rich and complex human language expressions with a small and limited number of rules, which is hard to extend. The modest parameter size and relatively simple model structure of LSTM-based methods are also not sufficient to support its ability to achieve a considerable level of language comprehension. In recent years, pre-trained language models have exhibited outstanding performances across many natural language processing tasks, attributed to their remarkable semantic understanding capabilities. Liang et al. (2022) and Akkasi and Moens (2021) utilized pre-trained language models to learn the semantic representation and achieved promising results. Akkasi and Moens (2021) also implemented and evaluated various techniques including a Multiview CNN (MVC), attention-based BiLSTM models, graph LSTM as well as a baseline rule-based system. Overall, compared to small-scale networks such as CNN, LSTM, and BiLSTM, pre-trained language models such as BERT (Kenton and Toutanova, 2019) and BioBERT (Lee et al., 2020) have shown better performances. However, these methods based on pre-trained language models are difficult to obtain enough semantic information solely in the context. Therefore, Zhang et al. (2023) introduced the external knowledge for BECRE and proposed the GECANet to fuse knowledge, which achieved state-of-the-art performance. In this paper, we put our emphasis on the entity relation path between biomedical events to capture richer cues for predicting event causal relations.

## 3 Method

Following the previous works (Zhang et al., 2023; Akkasi and Moens, 2021), we formulate BECRE as a classification problem. Given a sentence and an event pair, we predict whether there is a causal relation between the two events. Fig.2 schematically visualizes our approach, which mainly consists of three parts, including path reasoning module, representation module, and output module.

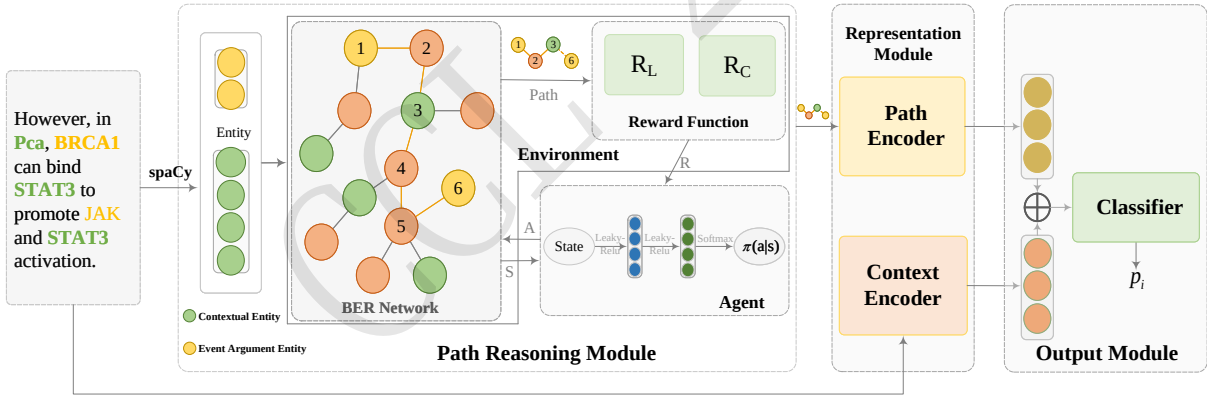



Figure 2: The illustration of our PRRN model, where "BER Network" is the abbreviation of biomedical entity relation network.

### 3.1 Path Reasoning Module

In this section, we introduce the process of reasoning paths based on reinforcement learning. Firstly, section 3.1.1 details the process of constructing a biomedical entity relation network. Secondly, for a given text and two events, we clarify the construction of reasoning tasks in section 3.1.2. Finally, section 3.1.3 introduces the reinforcement learning process.

#### 3.1.1 Biomedical Entity Relation Network Construction

In this paper, we construct a biomedical entity relation network from the Unified Medical Language System (UMLS) (Bodenreider, 2004) to reason optimal entity relation paths.

UMLS is a vast knowledge base which contains rich information about biomedical concepts and mainly includes three parts: Metathesaurus, Semantic Network and SPECIALIST Lexicon & Lexical Tools. Metathesaurus contains the information of biomedical concepts and each Metathesaurus concept is assigned a concepts unique identifier (CUI). Semantic Network consists of semantic types and semantic relationships. Semantic types (STY) are broad subject categories, like *Disease* or *Syndrome* or *Clinical Drug*. Semantic relationships are relationships that exist between semantic types. Every concept in Metathesaurus is assigned at least one semantic type.

The complete UMLs contains over five million terms, or names so far and the scale makes running a reasoning model on it costly. Therefore, a sub-network is constructed from it for reasoning in this paper. Specifically, MetaMap, an open source tool that can map biomedical text to the Metathesaurus, is employed to extract CUI of biomedical entities involved in datasets. Based on the set of CUI, we query 11-hop neighbor nodes of them from the Metathesaurus and save as a sub-network. Then the Semantic Network is introduced into this sub-network to further enrich it. The STY nodes in sub-network are connected with CUI nodes by the edge "CUI2STY". At last, we add reverse of every edge to obtain the final biomedical entity relation network.

As a result, the nodes in biomedical entity relation network are divided into CUI nodes and STY nodes, and the edges can be divided into those between CUI nodes, those between STY nodes, and those between CUI and STY. The final network consists of 8311 nodes, 34075 triples, and 85 relations. After obtaining the biomedical entity relation network, TransE (Bordes et al., 2013) is used to obtain $e^d \in \mathbb{R}^d$ and $r^d \in \mathbb{R}^d$, i.e. representations of nodes and relations in the network.

### 3.1.2 Reasoning Task Construction

Before performing reasoning, we construct reasoning tasks to prepare inputs of reasoning module. Specifically, a python library, spaCy[1] is employed to extract all biomedical entities in sentence to obtain $E_A = \{e|e \in BiomedicalEntity,\ e \in sentence\}$. And then we utilize the syntactic analysis tool of spaCy to analyze the syntactic dependency ancestor nodes of each entities. If the $k$-hop ancestor node of an entity is the trigger word of the head or tail event, and $k < K_{threshold}$ ($K_{threshold}$ is a hyper-parameter), the entity will be listed into the head entities set, $E_H$ or the tail entities set, $E_T$ , respectively. After obtaining $E_H$ and $E_T$, the inputs of reasoning module are constructed as $E_{sample}$ in Eq.(1).

$$E_{sample} = \{(e_H, e_T), \psi_M \mid e_H \in E_H, e_T \in E_T,\ e_H \neq e_T, \psi_M = E_A - \{e_H, e_T\}\}, \tag{1}$$

where $\psi_M$ denotes contextual entities set of size *M*, $(e_H, e_T)$ denotes the entity pair to be reasoned. We don't take the corresponding entity pair into consideration in our method when the $e_H$ is same as the $e_T$. For the convenience of locating entities in biomedical entity relation network in reasoning process, MetaMap mentioned in section 3.1.1 is employed again to map $E_{sample}$ to $E^{CUI}_{sample} = \{(e^{CUI}_H, e^{CUI}_T), \psi^{CUI}_M\}$.

### 3.1.3 Reinforcement Learning Process

The Reinforcement Learning (RL) process is typically modeled as an interaction between an agent and the external environment. In this paper, the process can be understood as that a policy-based agent reasons path hop by hop for input entity pairs. At each step, it choose action based on the current state of the environment to extend the path. The external environment evaluate the performance of agent with the reward function when the reasoning of single entity pair is completed. The agent will update its parameters based on evaluations of the environment, thus gradually approaching the optimal reasoning strategy encouraged by the reward function. We detail the RL process in PRRN below and present an overview of the process in Algorithm 1.

Firstly, for the input entities in $E^{CUI}_{sample}$, we locate them with CUI in biomedical entity relation network to obtain their node id, $E^{id}_{sample} = \{(e^{id}_H, e^{id}_T), \psi^{id}_M\}$ and TransE representations, $E^{d}_{sample} = \{(e^{d}_H, e^{d}_T), \psi^{d}_M\}$.

[1] https://spacy.io/

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1087-1098, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Secondly, at each step *t*, the agent will choose action based on the current state of the external environment. The action space in this paper is actually the relation set in biomedical entity relation network, denoted as $\xi = \{r_1, r_2, ..., r_{85}\}$. The state $s_t$ at step *t* is calculated as Eq.(2), where $e_t$ denotes the node where the agent is located at step *t*, $e_c$ denotes the pooling representation of the contextual entities and $||$ denotes the concatenation operation.

$$s_t = e_t^d \,||\, [e_T^d - e_t^d] \,||\, e_c^d \in \mathbb{R}^d,\ e_c^d = \frac{1}{M} \sum_{e \in \psi_M} e^d. \tag{2}$$

**Algorithm 1** Reinforcement Learning Process

restore parameter $\theta$ from supervised learning
**while** $i$<samples_num **do**
  $\{(e_H^{CUI}, e_T^{CUI}), \psi_M^{CUI}\}_i \xrightarrow{\text{match}} \{(e_H^{id}, e_T^{id}), \psi_M^{id}\}_i$
  $\{(e_H^{id}, e_T^{id}), \psi_M^{id}\}_i \xrightarrow{\text{query}} \{(e_H^{d}, e_T^{d}), \psi_M^{d}\}_i$
  $E_{sample}^d = \{(e_H^d, e_T^d), \psi_M^d\}$
  $\{(e_H^d, e_T^d), \psi_M^d\}_i \xrightarrow{Eq.(2)} s_0$
  **while** $t : 0 \to t_{max}$ **do**
    choose action, $s_t \xrightarrow{\pi(a|s_t,\theta)} r_t$
    state transition, $(r_t, s_t) \xrightarrow{Eq.(2)} s_{t+1}$
    **if** reach $e_T$ **then**
      break
    **end if**
  **end while**
  Path=$[(s_0, r_0), (s_1, r_1), \ldots, (s_n, *)]$
  **if** success **then**
    $R_A = \alpha \cdot R_L + (1 - \alpha) \cdot R_c$
  **else**
    $R_A = -1$
  **end if**
  update $\theta$ with Eq.(6)
**end while**

The agent in this paper is policy-based and implemented using a fully connected neural network, as the "Agent" part shown in Fig.2. The neural network is composed of two hidden layers and each layer is followed by a nonlinear activation function (i.e. LeakyRelu). The dimension of input layer and output layer is equal to the size of $s_t$ and $\xi$, respectively. It maps the current state to probability distribution across the entire action space and can be represented as a policy network $\pi_\theta(s, a) = p(a|s; \theta)$, where $\theta$ denotes parameters of the network. The agent choose most promising actions based on the probability distribution on the action space at each step *t*.

The agent will randomly transfer to the next node based on the transition probability matrix (denoted as Eq.(3) in the environment after choosing the action.

$$P\left(s_{t+1} = s' \mid s_t = s, a_t = r_i\right) = \begin{cases} \frac{1}{|\varphi_{r_i}^s|} & , s' \in \varphi_{r_i}^s, \\ 0 & , others, \end{cases} \tag{3}$$

where $\varphi_{r_i}^s = \{s' \mid r_i \in r_{s,s'}\}$ denotes the neighbor nodes of $s$ with relation $r_i$ as the edge. In this transition strategy, if the chosen action $r_i$ leads to an empty target state, the agent will go back to the original state $s_t$ and make another new decision.

Repeat the above decision-transition process, the path will continuously extend until the agent reaches the tail entity or the number of reasoning steps reaches the threshold $t_{max}$. When the reasoning of single

CCL 2024

entity pair is completed, the external environment will evaluate reasoning performance of the agent based on the reward function.

Given the reasoned path $p = \{e_0, r_0, e_1, r_1, ..., r_{L-1}, e_L\}$ with a length of $L$, the reward function in PRRN is designed for better capturing cues for predicting event causal relations and includes two components:

**Length-optimized Reward.** We set this reward component to encourage the agent to reach the target node with as few hops as possible. On the one hand, we observe that the shorter the path, the stronger the relation between the head and tail node it can express. Then it would be very beneficial for us to study the causal relation between events in BECRE task. On the other hand, shorter paths can also save computational resources and time and then improve reasoning efficiency. In addition, Length-optimized Reward are also aimed at mitigating the inherent challenges of sample inefficiency and high variance (Guo et al., 2021; Han et al., 2023) in policy-based RL algorithms when the path is longer. The length-optimized reward is defined as $R_L = \frac{1}{L}$.

**Contextual relevance reward.** This reward component encourages the agent to be closer to contextual entities during the reasoning process. The contextual entities, head entities, and tail entities share the same context and have strong correlation with each other. Therefore, using contextual entities information can quickly connect the head and tail entities, working together with length-optimized reward to improve reasoning efficiency. In addition, when the nodes on the path are closer to the contextual entity, it can better discover the cues hidden in the context about event causality, avoiding introducing noise that completely unrelated to the context. The contextual relevance reward is defined as $R_C$ in Eq.(4).

$$R_C = \frac{1}{L}\sum_{i=0}^{L} \cos\left(e_i^d, e_c^d\right),\ e_i \in p. \tag{4}$$

Finally, the total reward $R_A$ is obtained by weighting the above two reward components, as shown in the Eq.(5).

$$R_A = \alpha \cdot R_L + (1-\alpha) \cdot R_C, \tag{5}$$

where $\alpha$ is a hyper-parameter to balance two reward components. By weighting and integrating these two reward components, the agent can be encouraged to pursue both shorter path lengths and higher relevance to the context. In addition, when the agent fails to reach the target node after a certain round of interaction with the environment, we deem the reasoning as failed, and set $R_A = -1$ to give the agent a negative reward.

We update the policy network with Monte-Carlo Policy Gradient (Williams, 1992) and gradient descent, as shown in Eq.(6).

$$\begin{aligned} G_r(\theta) &= E_{\tau\sim\pi_\theta} \sum_{t=0}^{L} [-R(\tau)\nabla_\theta \log \pi(a_t|s_t;\theta)] \\ &\approx \nabla_\theta \sum_{t=0}^{L} [-\log \pi(a = r_t|s_t;\theta) R_A]. \end{aligned} \tag{6}$$

### 3.2 Representation Module

#### 3.2.1 Path Encoder

Recently, recurrent neural networks have been widely used in processing sequence data such as path information. Compared to vanilla GRU, BiGRU brings more comprehensive understanding of the information in the sequence. For the reasoned entity relation path $p = \{e_0, r_0, e_1, ..., r_{L-1}, e_L\}$, the BiGRU (Chung et al., 2014) is assigned to encode it in PRRN.

We initialize the representation of the path using TransE representations of edges and nodes. These node and relation representations integrate the topological information of nodes and relations in the biomedical entity relation network, allowing us to more thoroughly utilize the background knowledge of the knowledge base. Then the representation of entity relation paths is fed into BiGRU. We concatenate

CCL 2024

the last hidden state of BiGRU output from two directions to obtain the final representation of the entity relation path, as shown in Eq.(7).

$$V_p = BiGRU^D\left(\left[e_0^d, r_0^d, e_1^d, r_1^d, \ldots, r_{L-1}^d, e_L^d\right]\right) \in \mathbb{R}^D. \tag{7}$$

### 3.2.2 Context Encoder

Trigger words are important component of an event, and the semantic representation of trigger words is valuable reference for the model when predicting the relation between events. Therefore, we use BioBERT (Lee et al., 2020) to encode the text and extract the word representations corresponding to the trigger words. If the trigger words are more than one word, we extract the uniformed pooling representations of multiple trigger words, as shown in Eq.(8) and Eq.(9).

$$V_{head\_event} = \frac{1}{h_n}\sum_{i=1}^{h_n} f_c(head\_trigger_i|context) \in \mathbb{R}^D, \tag{8}$$

$$V_{tail\_event} = \frac{1}{t_n}\sum_{i=1}^{t_n} f_c(tail\_trigger_i|context) \in \mathbb{R}^D, \tag{9}$$

where $h_n$ and $t_n$ respectively denotes the number of trigger words for the head and tail events, and $f_c(w|context)$ denotes the contextual semantic representation of word $w$ from BioBERT (Lee et al., 2020). Benefiting from the attention mechanism (Bordes et al., 2013) of BioBERT (Lee et al., 2020), these two word representations simultaneously contain the semantic and contextual information of trigger words, which is effective in predicting causal relations between events.

## 3.3 Output Module

In section 3.2, we calculate the representation of entity relation paths and the contextual semantic information of events, respectively. For these two representations, we add the representation vectors of entity relation paths to the representation vectors of head and tail events, respectively. The two fusion vectors is then concatenated to calculate the final representation of each event pair, as shown in Eq.(10).

$$V_{feature} = [V_{head\_event} + V_p] \,||\, [V_{tail\_event} + V_p]. \tag{10}$$

And then $V_{feature}$ is fed into a fully connected neural network followed by a softmax function to obtain the final prediction. As shown in Eq.(11), $W_o$ denotes the parameters of the fully connected neural network, and $p_i$ denotes the predicted probability of *i*-th class.

$$p_i = softmax(W_o \cdot V_{feature}),\ W_o \in \mathbb{R}^{d_{class} \times D}. \tag{11}$$

## 3.4 Training Process

Following the previous work (Xiong et al., 2017; Silver et al., 2016), the training of policy network in agent is divided into two stages in our method including supervised learning and retraining with reinforcement learning. Supervised learning enables the policy network to have initial reasoning ability in a vast action space, thereby accelerating the convergence speed of reinforcement learning.

In the supervised learning stage, we run the breadth-first searching algorithm on the biomedical entity relation network to search paths between the input entity pair. The searched path is employed as the supervised information to update policy network. We use the policy gradient and set the reward to 1 to update the policy network in the agent. The gradient is shown in the Eq.(12).

$$G_s(\theta) = \bigtriangledown_\theta \sum_{t=0}^{n} log\pi(a = r_t|s_t;\theta), \tag{12}$$

where $s_t$ and $r_t$ are the states and actions of the *t*-th hop in the path, respectively. As for the sample set used in supervised learning, we conduct the reasoning sample construction as mentioned in section 3.1.2 and extract a sample set of $E^{CUI}_{sample}$ with size *samples_num*.

CCL 2024

Then in the reinforcement learning stage, we use the reward function to adjust the policy network. Firstly, the agent uses the prior knowledge obtained from the supervised learning stage to reason entity relation paths hop by hop. Next, the external environment will evaluate the performance of the agent using the reward function and feedback the evaluation to the agent. Finally, the policy network calculate the policy gradient based on this evaluation to update its parameters. By repeating this process, the policy network will gradually approach the optimal reasoning policy encouraged by the reward function.

We train the context encoder, path encoder and classifier together using cross entropy loss. In this stage, the policy network is called offline and does not participate in training. The loss we applied is shown in Eq.(13).

$$L = -\frac{1}{N}\sum_{i}\sum_{c=1}^{3}\mathbb{1}_c(i)log(p_{i,c}), \tag{13}$$

where $N$ denotes the number of samples, $\mathbb{1}_c(i)$ is an indicator function that is assigned 1 when $i = c$ and 0 when $i \neq c$ and $p_{i,c}$ denotes the predicted probability that the $i$-th sample belongs to $c$-th category.

## 4 Experiment

### 4.1 Experimental Setup

**Datasets and Metrics.** We evaluate our proposed method on two datasets, including Hpowell (Hahn-Powell et al., 2016) and BioCause (Mihăilă et al., 2013). Following prior works (Zhang et al., 2023; Akkasi and Moens, 2021), in Hpowell and BioCause, we only consider two types of relations, "E1 precedes E2" and "E2 precedes E1" as target relations and consider all the others as no relation, denoted as "None". Tab.1 presents the number of samples for each category in two datasets. We adopt Recall (R), Precision (P), and F1-score ($F_1$) as evaluation metrics.

| Dataset | Class | Number |
|---|---|---|
| Hpowell (Hahn-Powell et al., 2016) | E1 precedes E2 | 163 |
| | E2 precedes E1 | 27 |
| | None | 637 |
| Biocause (Mihăilă et al., 2013) | E1 precedes E2 | 36 |
| | E2 precedes E1 | 14 |
| | None | 3140 |

Table 1: Statistics of datasets used for the experiments.

**Implemention Details.** In the experiments, we use BioBERT-large as the context encoder. The representation dimension of token is 768. The $K_{threhold}$ mentioned in section 3.1.2 is set to 2, and $t_{max}$ mentioned in section 3.1.3 is set to 50. We train the context encoder, path Encoder and classifier with the learning rate=5e-5 and the dropout rate=0.1. Following the previous work (Zhang et al., 2023; Akkasi and Moens, 2021), we adopt the oversampling method to alleviate the sample imbalance problem in the Hpowell and BioCause datasets. Specifically, the categories "E2 precedes E1" and "E1 precedes E2" are sampled to equal numbers, matching the quantity of samples in the "None" category. The dimension of hidden layers in the policy network of Path Reasoning Module are 512 and 1024, respectively. We train TransE model on our biomedical entity relation network for 1000 training rounds[2].

**Baseline Methods.** We compare the proposed PRRN with the following baseline models. Akkasi and Moens. (2021) overviewed the BECRE task and contributed some credible benchmarks, as follows: **Rule-based Model** manually summarizes some causative verbs from the corpus and designs rules to predict the event causal relation based on causative verbs. **Graph state LSTM** models the document as a graph and captures a variety of dependencies among the input words which are represented by nodes of the graph. **Multiview-CNN** integrates feature representations of context calculated from three different sizes of convolution kernels. **BiLSTM-ATTENTION** utilizes BiLSTM to calculate the word representation of tokens and obtain the representation of texts by fusing representation of each token with attention mechanism. **ELMO-LSTM** simply applies ELMO as a word representation layer, and feeds representation result into a LSTM module. **BioBERT-MVC** uses four BioBERT (denoted as BioBERT1, BioBERT2, BioBERT3 and BioBERT4 below) that pre-trained on four different public

[2]The implemention we used can be found at https://www.github.com/thunlp/Fast-TransX

biomedical repositories to obtain the representation of each token, respectively and employs Multiview-CNN as the inference layer. In addition, **GECANet** (Zhang et al., 2023) introduces knowledge triplets from an external knowledge base and calculates the edge representation, node representation, and graph representation of the triplet set separately. Then it integrates them with context representation to predict event causal relations. For the above methods, Akkasi and Moens (2021) and Zhang et al. (2023) have both detailed the experimental results utilizing both raw data and oversampled data. Within this section, our focus is solely on comparing these methods under the oversampling strategy.

### 4.2 Experimental Results

| Source | Model | Hpowell | | | BioCause | | |
|---|---|---|---|---|---|---|---|
| | | P(%) | R(%) | $F_1$(%) | P(%) | R(%) | $F_1$(%) |
| Akkasi and Moens (2021) | Rule-based Model | - | - | - | 33.0 | 21.0 | 25.0 |
| | Graph state LSTM | 57.0 | 45.0 | 50.0 | 54.0 | 36.0 | 43.0 |
| | Multiview-CNN | 60.0 | 44.0 | 52.0 | 77.0 | 20.0 | 32.0 |
| | BiLSTM-ATTENTION | 54.0 | 49.0 | 52.0 | 59.0 | 35.0 | 44.0 |
| | ELMO-LSTM | 55.0 | 52.0 | 54.0 | 27.0 | 24.0 | 26.0 |
| | BioBERT1-MVC | 65.0 | 53.0 | 58.0 | 69.0 | 18.0 | 29.0 |
| | BioBERT2-MVC | 59.0 | 54.0 | 57.0 | 67.0 | 20.0 | 31.0 |
| | BioBERT3-MVC | 63.0 | 53.0 | 58.0 | 70.0 | 29.0 | 41.0 |
| | BioBERT4-MVC | 60.0 | 51.0 | 55.0 | **79.0** | 31.0 | 44.0 |
| Zhang et al. (2023) | GECANet | 64.6 | 60.5 | 62.5 | 78.2 | 32.0 | 45.5 |
| ours | PRRN | **65.7** | **61.5** | **63.6** | 75.8 | **44.0** | **55.7** |

Table 2: Experimental results on Hpowell and BioCause, where best results are bolded, and suboptimal results are underlined.

Tab.2 shows the results on the Hpowell and BioCause datasets, respectively. Overall, our method outperforms the methods proposed in previous works on both datasets in terms of F1-score. It can be seen that pre-trained language model based method like BioBERT3-MVC significantly outperforms Multiview-CNN model by 6.0% on Hpowell dataset. This is because the quite large parameter scale and the process of pre-training on large-scale corpora offer pre-trained language model stronger semantic representation ability than the small scale network. However, the method of solely utilizing the semantic representation ability of pre-trained language models ignores the causal cues stemmed from biomedical entities interactions. Therefore, the results of BioBERT3-MVC method on two datasets are 5.6% and 11.7% lower than our method, respectively, highlighting the effectiveness of learning to capturing deep cues with entity relation paths. In addition, our method surpasses the GECANet by 1.1% and 10.2% on the two datasets, respectively. This indicates that instead of directly incorporating external knowledge, delving into the entity relation path between biomedical events to discover the causal cues is a more beneficial method for BECRE.

Furthermore, in terms of recall and precision, our approach exhibits substantial enhancements over previous methods on the Hpowell dataset. On the BioCause dataset, while our method falls slightly behind Multiview-CNN, BioBERT-MVC and GECANet in precision, 1.2%, 3.2% and 2.4%, respectively, it achieves a remarkable enhancement in recall, which ensuring the reliability of our proposed method.

### 4.3 Ablation Study

In this section, we aim to ablate the major components in our model and evaluate the performance of the remaining model to understand contribution of each component. We examine the following ablated models: (1) "$-PRRN_{w/o.R_L}$" and "-$PRRN_{w/o.R_C}$" exclude the reward components $R_L$ and $R_C$ (respectively) from the overall reward when training the agent. (2) "-$PRRN_{w/o.path}$" indicates the model predicts event causal relations without the entity relation path. (3) "-$PRRN_{w/o.RL}$" indicates the model predicts event causal relations with the entity relation path obtained from the breadth-first searching algorithm.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1087-1098, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Tab.3 reports the performance of ablated models on the Hpowell dataset. As can be seen, removing the entity relation path or any reward components significantly hurts the overall performance of the model. The largest performance drop is due to the elimination of the entity relation path, suggesting that the entity relation path is critical to reveal the cues for predicting event causal relations. In addition, compared to the breadth-first searching algorithm, the superior performance of reinforcement learning clearly highlights the advantage to reason the entity relation path with reinforcement learning for BECRE.

| Method | P(%) | R(%) | $F_1$(%) | Δ |
|---|---|---|---|---|
| -PRRN$_{w/o.path}$ | 60.2 | 57.4 | 58.8 | - |
| -PRRN$_{w/o.R_L}$ | 65.9 | 59.0 | 62.2 | +3.4 |
| -PRRN$_{w/o.R_C}$ | 66.9 | 56.3 | 61.1 | +2.3 |
| -PRRN$_{w/o.RL}$ | **67.3** | 57.4 | 61.9 | +3.1 |
| PRRN(ours) | 65.7 | **61.6** | **63.6** | **+4.7** |

Table 3: Results of ablation study on reinforcement learning. Δ denotes improvements of $F_1$ compared to -PRRN$_{w/o.path}$.

### 4.4 Entity Relation Path Encoding Strategy Analysis

| Model | P(%) | R(%) | $F_1$(%) |
|---|---|---|---|
| LSTM | 61.8 | 57.9 | 59.8 |
| BiLSTM | 64.3 | 60.5 | 62.3 |
| GRU | **68.1** | 57.4 | 62.6 |
| BiGRU | 65.7 | **61.6** | **63.6** |

Table 4: Experimental results using different entity relation path encoders on the Hpowell dataset.

| Method | Hpowell | BioCause |
|---|---|---|
| ERP&Node | 61.5(↓ 2.2) | 52.1(↓ 3.6) |
| ERP&Edge | 61.1(↓ 2.6) | 53.9(↓ 1.8) |
| ERP&Full | **63.6** | **55.7** |

Table 5: Experimental results ($F_1$) of retaining only edges and nodes in ERP(Entity Relation Path) separately. ↓ means the decrease compared to ERP&Full.

**Effect of Entity Relation Path Encoder.** To analyze the effect of different entity relation path encoders on experimental results, we utilize four different path encoders, LSTM, BiLSTM, GRU, and BiGRU. The results are shown in Tab.4. From the results, we can observe that the best performing path encoders is BiGRU. This suggest that BiGRU has the capability to utilize the sequence information in the entity relation path more effectively, enabling more accurate aggregation of information from edges and nodes. In addition, the results also indicates that adopting bidirectional encoding for the entity relation path is more effective than unidirectional encoding.

**Effect of Nodes and Edges in the entity relation path.** To better know the effect of different path encoding strategies, we examine another two encoding strategies, "ERP&Node" and "ERP&Edge", which keeps only nodes and edges respectively in the paths in the phase of encoding. The results are shown in Tab.5, compared to using both edges and nodes simultaneously, eliminating each of them showed a reduced effect, indicating that both edge sequences and node sequences contain causal information between biomedical events. However, the complete path including both edges and nodes is more effective in predicting event causal relations.

### 4.5 The effect of hyper-parameter $\alpha$

In Path Reasoning Module, the hyper-parameter $\alpha$ is a adjustment term in Eq.(4), controlling the attention of the agent to length and contextual relevance of paths, respectively. To analyze the impact of $\alpha$ on model performance, we set $\alpha$ to different values to observe the $F_1$.

As shown in Fig.3, the $F_1$ score peaks when $\alpha = 0.5$, indicating that giving the same weight to the contextual relevance and length of the path can yield the best benefits on the Hpowell dataset. Moreover, there is a

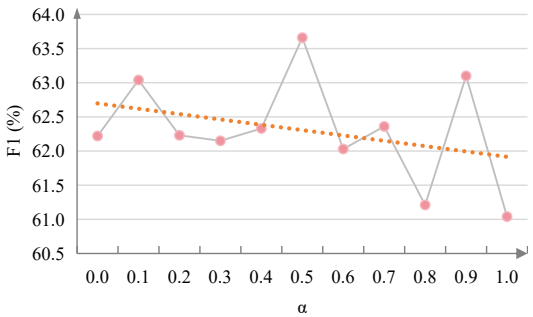



Figure 3: The effect of hyper-parameter $\alpha$

CCL 2024

discernible downward trend in the trendline, demonstrating that contextual relevance is more important than length of the path for capturing causal cues.

## Conclusion

In this paper, we emphasize the importance of the entity relation path for predicting event causal relations between biomedical events and present a novel model, namely, PRRN for Biomedical Event Causal Relation Extraction. For obtaining the valuable entity relation path, we propose a path reasoning module based on reinforcement learning and design length-optimized reward and contextual relevance reward to evaluate the performance of reinforcement learning agent. Our extensive experiments show that the PRRN outperforms the previous state-of-the-art model, demonstrating the advantage of the entity relation path reasoned under the control of two reward components. In the future, we will extend our proposed method to other related tasks in events relation extraction (e.g., for event causal relation extraction in universal domain).

## Acknowledgements

This work is supported by grant from the National Natural Science Foundation of China (No. 62076048).

## References

Abbas Akkasi and Mari-Francine Moens. 2021. Causal relationship extraction from biomedical text using deep neural models: A comprehensive survey. *Journal of biomedical informatics*, 119:103820.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26.

Hian Cañizares-Díaz, Alejandro Piad-Morffis, Suilan Estevez-Velarde, Yoan Gutiérrez, Yudivián Almeida Cruz, Andres Montoyo, and Rafael Muñoz. 2021. Active learning for assisted corpus construction: A case study in knowledge discovery from biomedical text. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 216–225.

Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*.

Hanlin Guo, Bowen Tan, Zhengzhong Liu, Eric P Xing, and Zhiting Hu. 2021. Text generation with efficient (soft) q-learning. *arXiv*.

Gus Hahn-Powell, Dane Bell, Marco A Valenzuela-Escárcega, and Mihai Surdeanu. 2016. This before that: Causal precedence in the biomedical domain. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pages 146–155.

Jiale Han, Mingxiao Feng, Wengang Zhou, and Houqiang Li. 2023. Sample efficient reinforcement learning with double importance sampling weight clipping. In *2023 IEEE Conference on Games (CoG)*, pages 1–8. IEEE.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Zhengzhong Liang, Enrique Noriega-Atala, Clayton Morrison, and Mihai Surdeanu. 2022. Low resource causal event detection from biomedical literature. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 252–263.

Claudiu Mihăilă and Sophia Ananiadou. 2014. Semi-supervised learning of causal relations in biomedical scientific discourse. *Biomedical engineering online*, 13:1–24.

Claudiu Mihăilă, Tomoko Ohta, Sampo Pyysalo, and Sophia Ananiadou. 2013. Biocause: Annotating and analysing causality in the biomedical domain. *BMC Bioinformatics*, 14:2.

Emmanuel Mutabazi, Jianjun Ni, Guangyi Tang, and Weidong Cao. 2021. A review on medical textual question answering systems based on deep learning approaches. *Applied Sciences*, 11(12):5456.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8:229–256.

Wenhan Xiong, Thien Hoang, and William Yang Wang. 2017. Deeppath: A reinforcement learning method for knowledge graph reasoning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Beibei Zhang, Lishuang Li, Dingxin Song, and Yang Zhao. 2023. Biomedical event causal relation extraction based on a knowledge-guided hierarchical graph network. *Soft Computing*, 27(22):17369–17386.

CCL 2024

# Joint Entity and Relation Extraction Based on Bidirectional Update and Long-Term Memory Gate Mechanism

**Yili Qian** * **Enlong Ren** **Haonan Xu**
School of Computer and Information Technology,
Shanxi University,Shanxi 030006
qyl@sxu.edu.cn

## Abstract

Joint entity recognition and relation extraction are important tasks in natural language processing. While some previous work has recognized the importance of relation information in joint extraction, excessively focusing on relation information without utilizing entity information may lead to information loss and affect the identification of relation tuples. Additionally, ignoring the utilization of original information may result in the loss of hierarchical and semantic information, further reducing the richness of information.To address these issues, we propose a bidirectional information updating mechanism that integrates entity and relation information, iteratively fusing fine-grained information about entities and relations. We introduce a long-term memory gate mechanism to update and utilize original information using feature information, thereby enhancing the model's ability for entity recognition and relation extraction. We evaluated our approach on two Chinese datasets and achieved state-of-the-art results.

## 1 Introduction

Joint entity recognition and relation extraction aim to extract entities and their relationships from unstructured text. The two common methods of relation extraction are pipeline extraction and joint extraction. Pipeline extraction (Mintz et al., 2009; Riedel et al., 2010; Chan and Roth, 2011) was a commonly used method in the early stages of relation extraction research. It divides relation extraction into two steps: first identifying entities, and then determining if there is a relationship between the identified entities. Due to the lack of interaction between the two steps, error propagation can occur. Subsequently, scholars began to research on joint extraction models (Miwa and Sasaki, 2014; Zhong and Chen, 2020; Li et al., 2022b) to address this issue. Some researchers have approached the problem from the perspective of labeling, designing appropriate labels for each word in a sentence to achieve end-to-end joint extraction (Zheng et al., 2017). Models of this kind are referred to as entity and relation extraction models based on joint decoding. However, in real-world scenarios, relation triplets are not simply one-to-one correspondences, different relations may involve entity overlaps. Joint decoding methods, due to limitations in labeling, cannot handle such complex relation triplets.

To address the extraction challenges posed by complex relationships, Zeng et al. (2018) and Zeng et al. (2020) first extract relationships, and then sequentially extract head entities and tail entities based on the extracted relationships, effectively avoiding relationship omissions. Liu et al. (2022b) divide joint extraction into two linked subtasks: the first subtask extracts head entities, and The second subtask is to extract relationships and tail entities based on the extracted head entities. This is achieved by mapping the head entity domain to the relationship and tail entity domains to enhance the model's ability to extract complex relationships. Later, researchers realized the importance of relational information and strengthened the extraction capabilities of models by integrating and utilizing relational information in various ways. Zhao et al. (2021) introduces relationship information as prior knowledge and combines it with extracted head entities to extract relationships and tail entities. Gao et al. (2022) first identifies the

* The corresponding author.
©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1099-1111, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

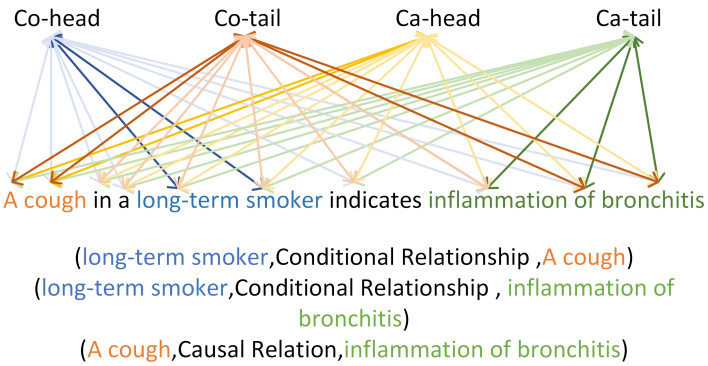



Figure 1: Example of Bidirectional Information Update. In the figure, the darker the arrow color, the greater the correlation between the word vectors and the dual-feature vectors. This shows that the richer the relation and entity information fused by the word vector. The dual-feature vectors also incorporate a wealth of word vector information, enhancing the representations of both.

relationships existing in the sentence, then randomly selects one relationship to merge into the sentence features to guide entity recognition. In order to reduce redundant information caused by complex networks, some methods (Liu et al., 2022a) use the same encoder to encode both relationship information and sentences, and achieve the fusion of the two parts of information through direct addition.

Despite the good results achieved by the aforementioned work, the following issues still exist:

- Lack of utilization of entity information. Xu et al. (2021) and Yuan et al. (2021) argue that relationships play a guiding role in triplet extraction and use attention mechanisms to update the contribution of each word in the sentence to the relationship. However, the lack of entity information leads to imbalance in word information, which is detrimental to the recognition of head/tail entities, thereby affecting the extraction of relationship tuples.

- Ignoring the original information leads to information loss. With the training of deep learning neural networks, although vector information is updated and new feature information is integrated, the original information is also altered as a result. Lack of original context and semantic information can reduce the accuracy of joint extraction.

To address the aforementioned issues, we propose a joint entity and relation extraction model based on bidirectional information updating and long-term memory gating mechanism. Firstly, we introduce entity information and relationship information as prior knowledge, represented together in a multi-dimensional vector, which we refer to as a dual-feature vector. Through an information updating mechanism, we sequentially update word vectors and dual-feature vectors. Multiple rounds of interactive updates allow for deep fusion of word feature vectors and dual-feature vectors. As shown in the Figure 1, we integrate specific relationships with head entity and tail entity information. ”Co-head” represents the head entity vector of the ”Conditional Relationship”, while ”Co-tail” represents the tail entity vector of the ”Conditional Relationship”. Similarly, ”Ca-head” represents the head entity vector of the ”Causal Relation”, and ”Ca-tail” represents the tail entity vector of the ”Causal Relation”, During the iterative fusion process with word vectors, the word vectors are enriched with rich relationship and entity information. The separate representation of head and tail entities also provides more precise entity information to the word vectors. Word vectors related to predefined information obtain enhanced representations, which are beneficial for further entity recognition and relationship extraction. Additionally, the pre-embedded dual-feature vectors also receive enhanced representations during the updating process. Subsequently, a relationship filtering module is employed to select the dual-feature vectors with the highest relevance, combining them with word feature vectors to guide the update of original word feature vectors. Finally, the combined feature vectors containing dual-feature information and original context information serve as the final word feature vectors for relation extraction.The main contributions of this work are as follows:

- We combine entity information with relationship information as prior knowledge and use an iterative information updating network to update word feature vectors, enriching them with rich entity and relationship information.
- We propose a feature-guided long-term memory gate mechanism to update the original word feature vectors containing contextual information, thereby preventing information loss.
- Our method achieves state-of-the-art results on two datasets.

## 2 Related Work

The traditional pipeline extraction approach (Cai et al., 2016; Zeng et al., 2014) sequentially extracts entities and relationships. Although flexible and simple, it faces two main issues. First, there is error propagation due to the independent nature of entity recognition and relationship extraction stages. Errors in entity recognition cannot be corrected during the relationship extraction stage, leading to irreversible impacts on the model's ability to accurately extract relationships. Second, there is a lack of information interaction between entity recognition and relationship extraction tasks, overlooking the inherent connection between entities and relationships. To address these issues, some scholars have proposed the concept of joint extraction to enhance the accuracy of extraction by strengthening the correlation between entity recognition and relationship extraction. Miwa and Bansal (2016) proposed an end-to-end joint extraction model where the authors used BiLSTM to extract contextual semantic information and utilized dependency syntax trees to extract features between entities. This approach achieved parameter sharing between the two subtasks, effectively alleviating the issue of error accumulation and improving extraction efficiency. However, it heavily relies on features generated by external tools and does not fully integrate the two subtasks.

Stanovsky et al. (2018) and Jia et al. (2022) utilize a sequence labeling approach to design labels, where relationship categories and entity information are designed as different labels. Each word is assigned an appropriate label and decoded based on these labels. However, since a word can only be assigned one label, the extraction efficiency of these models decreases as the complexity of relationships (such as entity overlap) increases. Takanobu et al. (2019) builds upon sequence labeling by treating entity information as parameters of the relationship. Through reinforcement learning, it enhances the interaction between entities and relationships, addressing the extraction of overlapping relationships. Liu et al. (2023a) uses a pointer network to label sentences, viewing the relationship extraction process as a mapping between two entities. It first extracts the subject in the sentence and then extracts the corresponding relationship and object based on the subject. This labeling approach solves the problem of weak expression capability in sequence labeling but may encounter label imbalance issues, requiring parameter tuning to address. Some researchers (Li and Fu, 2022; Nayak and Ng, 2020; Zeng et al., 2019) drew inspiration from machine translation, treating relation extraction as a sequence-to-sequence generation problem. They treat triplets as sequences to be translated, using a copying mechanism to find entities from the source sentence. Zhang et al. (2020) proposes the Seq2UMTree model, which addresses the label bias issue in sequence-to-sequence methods by shortening the decoding length and using tree-based decoding.

Treating relationship extraction as a table-filling problem (Wang and Lu, 2020; Kong and Xia, 2023) is also a commonly used method in joint extraction. Zhang et al. (2017), Ren et al. (2021) and Ning et al. (2023) generate a table for each relationship, labeling the start and end characters of the corresponding entities in the relationship table, and extract triplets using joint decoding. Li et al. (2022a) builds upon the table-filling approach by incorporating relationship information, strengthening global connections between same-position words in different relationship tables to achieve joint extraction.

## 3 Task Definition and Labeling Scheme

Given a sentence $X = [x_1, x_2, \ldots, x_m]$, where $x_i$ is the i-th word in the sentence, the objective is to extract triples $Y = \{y_1, y_2, \ldots, y_n\}$ that match predefined relation types within the sentence. Each $y = (h, r, t)$ represents a triple, where $h$ represents the head entity, $t$ represents the tail entity, and $r$ represents the relation type.In this paper, we adopt a labeling scheme similar to (Wang et al., 2020) and

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1099-1111, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

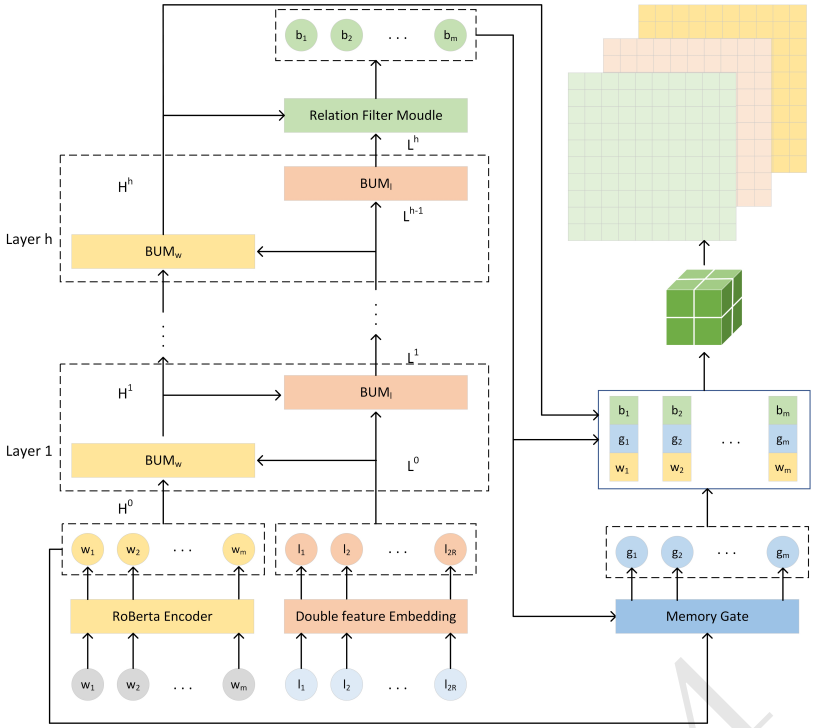



Figure 2: The overall framework of our proposed model.

design $2R+2$ labels based on the specific characteristics of the dataset. "N" represents that the text span $(w_i, w_j)$ does not form an entity, while "HT" represents that the text span $(w_i, w_j)$ forms an entity. For each type of relation, two labels are generated to annotate the head entity and tail entity, "HrH" represents that the start of the head entity of the r-th relation tuple is at position $w_i$, and the start of the tail entity is at position $w_j$. "TrT" represents that the end of the head entity of the r-th relation tuple is at position $w_i$, and the end of the tail entity is at position $w_j$. Both labels together form the head and tail entities of that relation. This labeling scheme for table filling generates $2R+1$ tables in total, with each relation assigned two tables labeled "HrH" and "TrT" respectively. A separate table is used solely to label which fields form entities, without distinguishing the type of entity.

## 4 Methodology

In this section, we will provide a detailed introduction to the joint entity and relation extraction model based on bidirectional updating and long-term memory gate mechanism. The framework we propose is illustrated in the Figure 2 and mainly consists of four parts:

(1) Encoding Module: Given a sentence, we encode the words in the sentence into vectors using a pre-trained model. Each relationship is combined with head/tail entities and represented using multi-dimensional vectors.

(2) Deep Fusion of Feature Information: We utilize multi-layer bidirectional updating networks to achieve deep fusion of word vectors and dual-feature vectors and ensure timely updates.

(3) Long-term Memory Gate Mechanism Guided by Most Relevant Dual-Feature Vectors: We first select the most relevant dual-feature vectors for each word in the sentence. Through a gate mechanism, we update the original word vectors to extract the original information of words, thereby avoiding the loss of semantic information.

(4) Relationship Extraction: We use the final word vectors for joint entity and relationship extraction.

### 4.1 Encoding Module

The pre-trained models trained on large-scale unlabeled text data have shown good performance in capturing context and semantics of sentences, and have been validated across various downstream tasks. For

CCL 2024

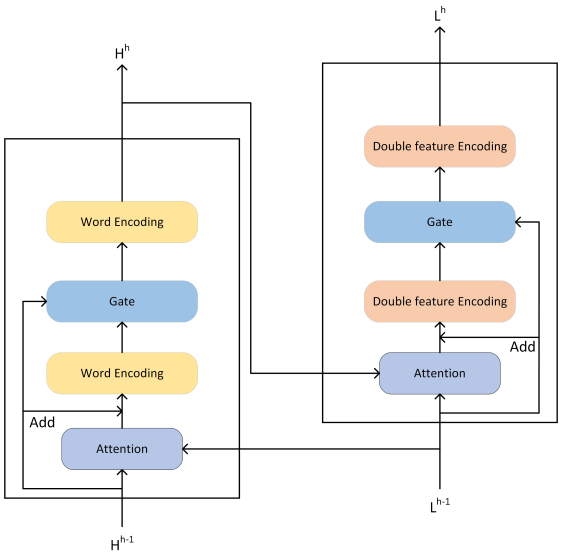



Figure 3: A complete layer of BUM.

a given sentence $X = [w_1, w_2, \ldots, w_m]$ , we obtain the contextual representation of the sentence by encoding it using Chinese-Roberta-Wwm-Ext-Large (Cui et al., 2021):

$$H = \text{Roberta}(X) \tag{1}$$

where $\mathrm{H} \in \mathrm{R}^{\mathrm{m}^* \mathrm{d}}$, $m$ as the number of characters in the sentence, and $d$ as the dimension of each word vector obtained after encoding. We initialize the predefined dual features as high-dimensional vectors and then obtain the vector representation of each feature through a linear layer mapping:

$$L = W_1 E + b_1 \tag{2}$$

where $E \in R^{2R^*d}$ is the dual-feature vector, $L \in R^{2R^*d}$, $R$ is the number of predefined relations, and $d$ is the dimensionality of each dual feature vector.

## 4.2 Deep Fusion of Feature Information

### 4.2.1 Vector Updating Mechanism

We have designed a Bidirectional Updating Mechanism (BUM) to acquire and update the semantic information between words and dual features. "Bidirectional" refers to the updates of dual feature vectors to word vectors and word vectors to feature vectors. The updating process for both follows the same mechanism, and we will only introduce one of them here. As shown in the Figure 3, for a given word vector $\{h_i\}_{i=1}^{m}$ and dual feature vector $\{l_j\}_{j=1}^{2R}$, we use an attention mechanism to perform semantic fusion and update between the two. We first use $w_i$ as the query, $l_j$ as the key and value to update the information of $w_i$, and then use residual connections to prevent gradient vanishing during training:

$$\begin{aligned} a_{ij} &= W_q w_i M W_k l_j^T \\ \mu_{ij} &= \frac{\exp\left(a_{ij}\right)}{\sum_{s \in \mathrm{N}_i} \exp\left(a_{is}\right)} \\ w_i' &= w_i + \sum_{j=1}^{2R} \mu_{ij} W_v l_j \end{aligned} \tag{3}$$

where $W_q, W_k, W_v, M \in R^{d^*d}$ are trainable weights, and $\mu_{ij}$ is the weights between $w_i \in R^d$ and $l_j \in R^d$.

We use a gate mechanism similar to Zhao et al. (2021) to control the information flow and maintain

CCL 2024

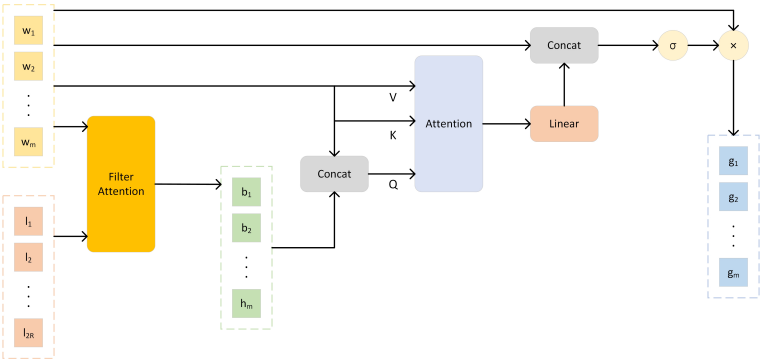



Figure 4: The detailed structure of the long-term memory gate mechanism.

nonlinear capabilities:

$$\begin{aligned} \theta_i &= \text{sigmoid}\left(W_z\left[w_i; w_i'\right]\right) \\ \tilde{w}_i &= \theta_i \odot w_i' + (1-\theta_i) \odot w_i \end{aligned} \tag{4}$$

where $W_z \in R^{d*d}$ is trainable weights, $\theta_i$ is the retention ratio of information, $\odot$ denotes element-wise multiplication, and $\tilde{w}_i$ is the vector of words updated through the bidirectional updating mechanism.

#### 4.2.2 Vector Iterative Fusion

We can summarize the aforementioned bidirectional update mechanism as follows:

$$\begin{aligned} \tilde{w}_i &= BUM_w\left(w_i, \{l_j\}_{j=1}^{2R}\right) \\ \tilde{l}_j &= BUM_l\left(l_j, \{w_i\}_{i=1}^{m}\right) \end{aligned} \tag{5}$$

where $BUM_w$ and $BUM_l$ are responsible for updating the bidirectional feature vectors to word vectors and word vectors to bidirectional feature vectors, respectively. Together, they form a complete unit called $BUM$. $w_i$ is the word vectors before the update, $l_j$ is the bidirectional feature vectors before the update, $\tilde{w}_i$ is the word vectors after the update, and $\tilde{l}_j$ is the bidirectional feature vectors after the update.

We achieve the iterative fusion of word vectors and dual feature vectors through multiple layers of $BUM$ units.Given the sentence context representation $H = [w_1, w_2, \ldots, w_m]$ and the dual feature vector representation $L = [l_1, l_2, \ldots, l_{2R}]$, we first update the word vectors and then use the updated word vectors to update the dual feature vectors. To prevent gradient vanishing, we introduce residual connections. Therefore, we represent the update process of the h-th layer as follows:

$$\begin{aligned} w_i^h &= w_i^{h-1} + BUM_w\left(w_i^{h-1}, \left\{l_j^{h-1}\right\}_{j=1}^{2R}\right) \\ l_j^h &= l_j^{h-1} + BUM_l\left(l_j^{h-1}, \left\{w_i^{h-1}\right\}_{i=1}^{m}\right) \end{aligned} \tag{6}$$

where $w_i^h$ is the update of the previous layer's output word vector by the h-th layer of $BUM_w$, and $l_j^h$ is the update of the previous layer's output double feature vector by the h-th layer of $BUM_l$.

### 4.3 Long-Term Memory Gate Mechanism Guided by Most Relevant Double Feature Vectors

In order to reduce the interference of redundant feature vectors, we designed a feature filtering module with the aim of finding the double feature vector that has the highest relevance with each word vector. Additionally, we developed a long-term memory gate mechanism guided by the most relevant double feature vectors to update and utilize the original word vectors, preventing the loss of original contextual information and achieving long-term memory effects. The detailed structure is illustrated in the Figure 4.

CCL 2024

#### 4.3.1 Double Feature Encoding and Filtering Module

We believe that not all double feature vectors will contribute positively to entity recognition and relation extraction. Indiscriminate use of all double feature vectors may introduce redundant information and consequently affect the accuracy of identification. Therefore, we employ an attention mechanism to determine the most relevant double feature vectors for each word and utilize them accordingly, The specific calculation process is as follows:

$$\begin{aligned} A_{i,j} &= \text{soft}\max\left(W_q w_i^h M W_k l_j'^T\right) \\ A'_{i,j} &= \begin{cases} 1, \text{ if } j = \text{argmax}\left(A_i\right) \\ 0, \text{ otherwise} \end{cases} \\ b_i &= \sum A'_{i,j} l'_j \end{aligned} \tag{7}$$

where $W_q, W_k, W_v, M \in R^{d*d}$ are the trainable weights, $A_{i,j}$ is the weights for the word vector and double feature vector, $A'_{i,j}$ is setting the position of the most relevant double feature vector for each word vector to 1 and others to 0, for subsequent computations. $b_i \in R^{1*d}$ is the most relevant double feature vector representation for each word vector.

#### 4.3.2 Long-Term Memory Gate Mechanism

After iterative fusion, the original context information is lost. We designed a gate mechanism based on feature encoding guidance to combine the filtered dual-feature vectors with the original word vectors, update the original word vectors, and then further filter the necessary information through the gate mechanism to reduce the loss of original information and achieve long-term memory. The specific implementation of this module is as follows:

$$\begin{aligned} w'_i &= W_c\left[W_a w_i; W_b b_i\right] \\ w''_i &= \text{soft}\max\left(W_q w'_i W_k l_j'^T\right) W_v l_j \\ w_i^{\text{gate}} &= \sigma\left(W_s\left[w'_i; w''_i\right]\right) w'_i \end{aligned} \tag{8}$$

where $W_a, W_b, W_c, W_k, W_k, W_v, W_s \in R^{d*d}$ are trainable parameters. $w'_i$ is the combined representation of the i-th original word vector and the filtered dual-feature vector. $[;]$ denotes the concatenation operation between the two vectors. $\sigma$ is the updated vector of the original word vector. $w_i^{\text{gate}}$ is the activation function, and H is the output vector from the gate mechanism.

### 4.4 Relation Extraction

We combine the final word vector output from the last layer of the bidirectional updating network, the word vector output from the long-term memory gate mechanism, and the most relevant dual-feature vectors from the feature filtering module to form the ultimate word vector. This ensures that the word vector contains entity and relationship information, context, and semantic information. To maximize the full interaction between each pair of words in the sentence, we obtain the final tabular feature vector using the following computational formula:

$$\begin{aligned} w_e &= \left[w_i^h; w_i^{\text{gate}}; l_i\right] \\ w_{i,j} &= W_r \sigma\left(w_{e,i} \circ w_{e,j}\right) \end{aligned} \tag{9}$$

where $w_e$ represents the final word vector, $\circ$ denotes the element-wise Hadamard product, and C is the joint vector representation of the i-th and j-th words. We obtain the labels for word pairs using the following formula:

$$\begin{aligned} P\left(z_{i,j}\right) &= \text{softmax}\left(W_p w_{i,j}\right) \\ \text{label}\left(w_i, w_j\right) &= \arg\max_t P\left(z_{i,j} = t\right) \end{aligned} \tag{10}$$

CCL 2024

where $P\left(z_{i,j}\right)$ is the probability of word pair $(w_i, w_j)$ being recognized as label $t$.

### 4.5 Loss Function

We use the cross-entropy loss function to calculate the difference between the predicted results and the true labels:

$$L(t, \tilde{t}) = -\sum_{c=1}^{2R+2} t_c \log \tilde{t}_c \tag{11}$$

where $t_c$ is the true label, $\tilde{t_c}$ is the predicted label.

## 5 Experiments

### 5.1 Datasets

We validate the performance of our model on two Chinese datasets CMedCausal (Li et al., 2022c) and CFinCausal [0]. CMedCausal is from the CHIP2022 academic evaluation task for medical causal entity relation extraction, The text is derived from online consultations and medical encyclopedia data from medical websites, with a length exceeding 200 Chinese characters. It includes causal relationships, conditional relationships, and hierarchical relationships. CFinCausal is a financial causal relation extraction dataset provided by the Tonghuashun platform in 2022, The dataset only contains causal relationships. The statistical results for the two datasets are shown in the Table 1. It is worth noting that in the CMedCausal dataset, the tail entity of the conditional relation tuple constitutes a complete causal relation tuple, We split the conditional relationships in Figure 1 into three relationship triplets. A complete conditional relation tuple is formed only when both the head entity and the tail entity of a causal relation are simultaneously related to another entity, forming a complete conditional relation tuple.

| Datasets | Train | Vaild | Test |
|---|---|---|---|
| CMedCausal | 899 | 100 | 1000 |
| CFinCausal | 1800 | 200 | 500 |

Table 1: Statistics of the data sets used in the experiment.

### 5.2 Baselines and Evaluation Metrics

To evaluate the model performance, we used the following powerful models as baselines: **Casrel** (Wei et al., 2019), **GPLinker** (Su, 2022), **TPLinker** (Wang et al., 2020), **GRTE** (Ren et al., 2021), **DEPT** (Liu et al., 2023b), **Onerel** (Shang et al., 2022), **UniRel** (Tang et al., 2022), **PRGC** (Zheng et al., 2021). We use $Macro\text{-}F1$ as the evaluation metric, and the specific calculation process is as follows:

$$Macro\text{-}F1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i + R_i} \tag{12}$$

where $P_i$ is the ratio of true predicted samples for the i-th relation to the predicted samples for the i-th relation, and $R_i$ is the ratio of true predicted samples for the i-th relation to the actual samples for the i-th relation.

### 5.3 Experimental Parameters

We set the hidden layer size of the pre-trained model to 1024. On the CMedCausal/CFinCausal datasets, we set the learning rate to 1e-4/5e-5, maximum character length of input sentences to 512/256, batch size for training to 8, and the dimension of the embedded double feature vectors to 1024. We employed dropout with a rate of 0.2 to prevent overfitting during training.

[0]The dataset can be obtained from http://contest.aicubes.cn/.

### 5.4 Experimental Results

Table 2 presents the results of our model and other baseline methods on the datasets. Our model outperforms other baseline methods on both datasets, with an improvement of 3.5% on the CMedCausal dataset and 1.4% on the CFinCausal dataset. The experimental results demonstrate the superiority of our approach. Casrel performs poorly on both datasets, likely due to its two-stage relation triplet extraction process, which is susceptible to error accumulation and puts it at a disadvantage compared to models that extract triplets simultaneously. Our approach and other baseline methods share the similarity of extracting triplets in one step, effectively avoiding the drawbacks of two-stage extraction models. The improvement in F1 score confirms the effectiveness of incorporating entity and relation information in our model, which enhances the efficiency of joint extraction.

| Method | CMedCausal | CFinCausal | | |
|---|---|---|---|---|
| | Macro-F1 | Prec. | Rec. | Macro-F1 |
| Casrel | 21.6 | 61.9 | 51.0 | 55.9 |
| PRGC | 32.8 | 61.9 | 53.1 | 57.2 |
| TPLinker | 33.2 | 57.0 | 56.8 | 56.9 |
| DEPT | 34.5 | 50.8 | 51.8 | 51.3 |
| Onerel | 35.9 | 56.9 | 58.0 | 57.4 |
| UniRel | 37.9 | 62.3 | 54.3 | 58.0 |
| GPLinker | 38.3 | **63.5** | 54.3 | 58.5 |
| GRTE | 38.4 | 57.4 | 58.9 | 58.1 |
| **Ours** | **41.9** | 60.5 | **59.3** | **59.9** |

Table 2: Results of different methods on CMedCausal and CFinCausal datasets.

| Method | Causal | | | Conditional | | | Hypernym | | |
|---|---|---|---|---|---|---|---|---|---|
| | Prec. | Rec. | F1 | Prec. | Rec. | F1 | Prec. | Rec. | F1 |
| Casrel | 28.3 | 26.9 | 27.6 | 3.0 | 1.7 | 2.2 | 41.4 | 30.3 | 35.0 |
| PRGC | 47.7 | 46.7 | 47.2 | 7.7 | 2.8 | 4.1 | 49.8 | 44.7 | 47.1 |
| TPLinker | 46.4 | 47.2 | 46.8 | 7.8 | 3.9 | 5.2 | 44.7 | 50.9 | 47.6 |
| DEPT | 48.5 | 46.2 | 47.3 | 10.4 | 3.9 | 5.7 | 53.1 | 48.0 | 50.4 |
| Onerel | 49.4 | 50.0 | 49.7 | 8.8 | 5.3 | 6.6 | 48.6 | 54.3 | 51.3 |
| UniRel | 53.7 | 54.3 | 54.0 | 6.2 | 3.9 | 4.8 | 52.7 | 57.3 | 54.9 |
| GPLinker | **58.3** | 49.8 | 53.7 | 11.8 | 3.9 | 5.9 | 58.1 | 52.6 | 55.2 |
| GRTE | 50.3 | **54.5** | 52.3 | 11.5 | **9.8** | 10.6 | 45.3 | **61.9** | 52.3 |
| **Ours** | 58.2 | 51.4 | **54.6** | **22.6** | 8.5 | **12.3** | **59.7** | 58.3 | **59.0** |

Table 3: Detailed results of the three relationships for different methods on the CMedCausal dataset.

### 5.5 Detailed Results of Different Relationships

Table 3 shows the detailed performance metrics for three relations in the CMedCausal dataset, where our proposed method outperforms others across these relations. Although this dataset does not categorize complex triplets explicitly, our decomposition of conditional relations implies that we can treat them as complex relation triplets. The performance metrics for this relation also indicate that it poses the highest extraction difficulty. Among the compared methods, GRTE leads in extracting this relation triplet with an $F1$ score of 10.6, while ours achieves an $F1$ score of 12.3, representing a 1.7% improvement, demonstrating our method's advantage in extracting complex relation triplets. GRTE achieved the best recall rates on all three types of relationships because this method does not separately label entities using a table. When decoding the relationship table, all possible triplets are extracted, which effectively reduces the omission of correct relationship tuples. However, this approach lacks individual entity judgment,

CCL 2024

which increases the probability of incorrect relationship tuples. Furthermore, our method exhibits the most significant lead in extracting conditional relations, surpassing the best baseline method by 3.8%.

### 5.6 Ablation study

To verify the effectiveness of each component, we conducted ablation experiments on the CMedCausal dataset, and the results are shown in Table 4. Removing either of the two gating mechanisms had a negative impact on the model performance, indicating the positive role of the gating mechanisms in the model. Substituting the updated word vectors with raw word vectors directly, which results in the absence of feature-guided semantic updates, demonstrated that this component not only utilizes the original contextual information but also effectively updates semantic information. Removing the feature filtering module led to a significant decrease in model performance, highlighting the importance of the most relevant features for subsequent tasks. Ceasing the update of feature vectors affected the utilization of updates for entity and relation information from word vectors, while halting the update of word vectors resulted in the worst model performance, severely impacting the deep fusion of feature information and word information. Missing entity and relation information in word vectors directly affected the update of feature vectors, indirectly affecting the subsequent feature filtering module and the utilization of original information.

| Method | Macro-F1 |
|---|---|
| Complete model | 41.9 |
| -word vector update | 38.9 |
| -feature filtering module | 39.4 |
| -feature vector update | 40.5 |
| -raw word vectors update | 40.8 |
| -memory gate mechanism | 41.4 |
| -gate mechanism in the update network | 41.6 |

Table 4: Abaltion experimental results on CMedCausal dataset.

### 5.7 The Number of Layers for Bidirectional Update Networks

The impact of different layers of bidirectional update networks on model performance is illustrated in Figure 5. The comparison between 0-layer and 1-layer networks shows a significant improvement in model performance, indicating the effectiveness of our proposed method. The 2-layer network shows a slight improvement over the 1-layer network, reaching optimal performance for the model, which suggests that as the number of update network layers increases, the model utilizes more entity and relation information. However, when the network reaches 3 layers, the model performance starts to decline. Therefore, we set the optimal number of layers for the model to be 2.

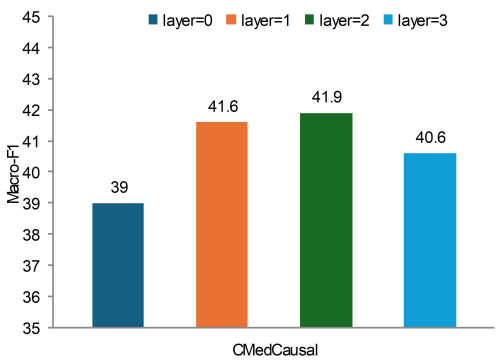



Figure 5: Results of different bidirectional update network layers on the CMedCausal dataset.

CCL 2024

## 6 Conclusion

In this study, we propose a novel joint extraction model that treats joint extraction as a table filling problem, incorporating the joint representation of entity and relation information as prior knowledge and iteratively updating and enriching the information of word vectors. We achieve the retention and utilization of original information through a long-term memory gate mechanism. Experimental results on two datasets demonstrate that our approach effectively integrates entity and relation information, outperforming existing strong baseline models. It also shows superiority in extracting complex relationships.

## Acknowledgements

This work was supported by the 1331 Engineering Project of Shanxi Province, China.

## References

Rui Cai, Xiaodong Zhang, and Houfeng Wang. 2016. Bidirectional recurrent convolutional neural network for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 756–765.

Yee Seng Chan and Dan Roth. 2011. Exploiting syntactico-semantic structures for relation extraction. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 551–560.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

Chen Gao, Xuan Zhang, Hui Liu, Wei Yun, and Jiahao Jiang. 2022. A joint extraction model of entities and relations based on relation decomposition. *International Journal of Machine Learning and Cybernetics*, 13(7):1833–1845.

Shengbin Jia, E Shijia, Ling Ding, Xiaojun Chen, and Yang Xiang. 2022. Hybrid neural tagging model for open relation extraction. *Expert Systems with Applications*, 200:116951.

Wenjun Kong and Yamei Xia. 2023. Care: Co-attention network for joint entity and relation extraction. *arXiv preprint arXiv:2308.12531*.

Zhe Li and Luoyi Fu. 2022. A relation-aware span-level transformer network for joint entity and relation extraction. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Hao Li, Md Tauhidul Islam, Kehan Huangliang, Zugang Chen, Kaiyang Zhao, and Haixian Zhang. 2022a. Setff: A semantic enhanced table filling framework for joint entity and relation extraction. In *Pacific Rim International Conference on Artificial Intelligence*, pages 169–182. Springer.

Lishuang Li, Zehao Wang, Xueyang Qin, and Hongbin Lu. 2022b. Dual interactive attention network for joint entity and relation extraction. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 259–271. Springer.

Z Li, M Chen, Z Ma, et al. 2022c. Cmedcausal: Chinese medical causal relationship extraction dataset. *J. Med. Inform*, 43(12):23–27.

Yang Liu, Jinpeng Hu, Xiang Wan, and Tsung-Hui Chang. 2022a. A simple yet effective relation information guided approach for few-shot relation extraction. *arXiv preprint arXiv:2205.09536*.

Yibo Liu, Feng Wen, Teng Zong, and Taowei Li. 2022b. Research on joint extraction method of entity and relation triples based on hierarchical cascade labeling. *IEEE Access*, 11:9789–9798.

PeiYu Liu, JunPing Du, YingXia Shao, and Zeli Guan. 2023a. Relation extraction model based on semantic enhancement mechanism. In *2023 5th International Conference on Data-driven Optimization of Complex Systems (DOCS)*, pages 1–7. IEEE.

Xu Liu, Ruiqi Cai, and Yonggang Zhang. 2023b. A 2d entity pair tagging scheme for relation triplet extraction. In *International Conference on Knowledge Science, Engineering and Management*, pages 16–29. Springer.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011.

Makoto Miwa and Mohit Bansal. 2016. End-to-end relation extraction using lstms on sequences and tree structures. *arXiv preprint arXiv:1601.00770*.

Makoto Miwa and Yutaka Sasaki. 2014. Modeling joint entity and relation extraction with table representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1858–1869.

Tapas Nayak and Hwee Tou Ng. 2020. Effective modeling of encoder-decoder architecture for joint entity and relation extraction. In *Proceedings of the AAAI conference on artificial intelligence*, number 05, pages 8528–8535.

Jinzhong Ning, Zhihao Yang, Yuanyuan Sun, Zhizheng Wang, and Hongfei Lin. 2023. Od-rte: A one-stage object detection framework for relational triple extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11120–11135.

Feiliang Ren, Longhui Zhang, Shujuan Yin, Xiaofeng Zhao, Shilei Liu, Bochao Li, and Yaduo Liu. 2021. A novel global feature-oriented relational triple extraction model based on table filling. *arXiv preprint arXiv:2109.06705*.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010, Proceedings, Part III 21*, pages 148–163. Springer.

Yu-Ming Shang, Heyan Huang, and Xianling Mao. 2022. Onerel: Joint entity and relation extraction with one module in one step. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 11285–11293.

Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and Ido Dagan. 2018. Supervised open information extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 885–895.

J Su. 2022. Gplinker: entity-relation joint extraction based on globalpointer.

Ryuichi Takanobu, Tianyang Zhang, Jiexi Liu, and Minlie Huang. 2019. A hierarchical framework for relation extraction with reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7072–7079.

Wei Tang, Benfeng Xu, Yuyue Zhao, Zhendong Mao, Yifeng Liu, Yong Liao, and Haiyong Xie. 2022. Unirel: Unified representation and interaction for joint relational triple extraction. *arXiv preprint arXiv:2211.09039*.

Jue Wang and Wei Lu. 2020. Two are better than one: Joint entity and relation extraction with table-sequence encoders. *arXiv preprint arXiv:2010.03851*.

Yucheng Wang, Bowen Yu, Yueyang Zhang, Tingwen Liu, Hongsong Zhu, and Limin Sun. 2020. Tplinker: Single-stage joint extraction of entities and relations through token pair linking. *arXiv preprint arXiv:2010.13415*.

Zhepei Wei, Jianlin Su, Yue Wang, Yuan Tian, and Yi Chang. 2019. A novel cascade binary tagging framework for relational triple extraction. *arXiv preprint arXiv:1909.03227*.

Wenqiang Xu, Shiqun Yin, Junfeng Zhao, and Ting Pu. 2021. Deep semantic fusion representation based on special mechanism of information transmission for joint entity-relation extraction. In *PRICAI 2021: Trends in Artificial Intelligence: 18th Pacific Rim International Conference on Artificial Intelligence, PRICAI 2021, Hanoi, Vietnam, November 8–12, 2021, Proceedings, Part II 18*, pages 73–85. Springer.

Yue Yuan, Xiaofei Zhou, Shirui Pan, Qiannan Zhu, Zeliang Song, and Li Guo. 2021. A relation-specific attention network for joint entity and relation extraction. In *International joint conference on artificial intelligence*. International Joint Conference on Artificial Intelligence.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers*, pages 2335–2344.

CCL 2024

Xiangrong Zeng, Daojian Zeng, Shizhu He, Kang Liu, and Jun Zhao. 2018. Extracting relational facts by an end-to-end neural model with copy mechanism. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 506–514.

Xiangrong Zeng, Shizhu He, Daojian Zeng, Kang Liu, Shengping Liu, and Jun Zhao. 2019. Learning the extraction order of multiple relational facts in a sentence with reinforcement learning. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 367–377.

Daojian Zeng, Haoran Zhang, and Qianying Liu. 2020. Copymtl: Copy mechanism for joint extraction of entities and relations with multi-task learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9507–9514.

Meishan Zhang, Yue Zhang, and Guohong Fu. 2017. End-to-end neural relation extraction with global optimization. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 1730–1740.

Ranran Haoran Zhang, Qianying Liu, Aysa Xuemo Fan, Heng Ji, Daojian Zeng, Fei Cheng, Daisuke Kawahara, and Sadao Kurohashi. 2020. Minimize exposure bias of seq2seq models in joint entity and relation extraction. *arXiv preprint arXiv:2009.07503*.

Kang Zhao, Hua Xu, Yue Cheng, Xiaoteng Li, and Kai Gao. 2021. Representation iterative fusion based on heterogeneous graph neural network for joint entity and relation extraction. *Knowledge-Based Systems*, 219:106888.

Suncong Zheng, Feng Wang, Hongyun Bao, Yuexing Hao, Peng Zhou, and Bo Xu. 2017. Joint extraction of entities and relations based on a novel tagging scheme. *arXiv preprint arXiv:1706.05075*.

Hengyi Zheng, Rui Wen, Xi Chen, Yifan Yang, Yunyan Zhang, Ziheng Zhang, Ningyu Zhang, Bin Qin, Ming Xu, and Yefeng Zheng. 2021. Prgc: Potential relation and global correspondence based joint relational triple extraction. *arXiv preprint arXiv:2106.09895*.

Zexuan Zhong and Danqi Chen. 2020. A frustratingly easy approach for entity and relation extraction. *arXiv preprint arXiv:2010.12812*.

CCL 2024

# MFE-NER: Multi-feature Fusion Embedding for Chinese Named Entity Recognition

**Jiatong Li** *
The Hong Kong Polytechnic University
jiatong.li@connect.polyu.hk

**Kui Meng**†
Shanghai Jiao Tong University
mengkui@sjtu.edu.cn

## Abstract

In Chinese Named Entity Recognition, character substitution is a complicated linguistic phenomenon. Some Chinese characters are quite similar as they share the same components or have similar pronunciations. People replace characters in a named entity with similar characters to generate a new collocation but refer to the same object. As a result, it always leads to unrecognizable or mislabeling errors in the NER task. In this paper, we propose a lightweight method, MFE-NER, which fuses glyph and phonetic features to help pre-trained language models handle the character substitution problem in the NER task with limited extra cost. Basically, in the glyph domain, we disassemble Chinese characters into Five-Stroke components to represent structure features. In the phonetic domain, an improved phonetic system is proposed in our work, making it reasonable to describe phonetic similarity among Chinese characters. Experiments demonstrate that our method performs especially well in detecting character substitutions while slightly improving the overall performance of Chinese NER.

## 1 Introduction

Recently, pre-trained language models have been widely used in Natural Language Processing (NLP) (Zhang et al., 2021), constantly refreshing the benchmarks of various NLP tasks (Li et al., 2024b; Li et al., 2024a). By applying the transformer structure, semantic features can be extracted more accurately. However, in the Named Entity Recognition (NER) area, tricky problems still exist. Most significantly, the character substitution problem severely affects the performance of NER models. To make things worse, this problem has become even more common these years, especially in social media, which leads to the dissemination of fake news (Frermann et al., 2023). Due to the particularity of Chinese characters, there are multiple ways to replace original Chinese characters in a word. Characters with similar meanings, shapes, or pronunciations can be selected for character substitution. A simple example shown in Figure 1 is a Chinese word, which represents a famous place in Shanghai. Here, all three characters in the word are substituted by other characters with similar glyphs or pronunciations. After substitution, this word looks more like the name of a person rather than a place.

In practice, it is extremely hard for those pre-trained language models to tackle this problem (Mao et al., 2022). Currently, the tasks for pre-training Chinese language models are mainly focused on the semantic domain, neglecting glyph and phonetic features (Li et al., 2020). However, most character substitution cases exist in glyph and phonetic domains. At the same time, hot topics on social media are changing rapidly, creating new expressions or substitutions for original words every day, while it is technically impossible for pre-trained models to include all possible collocations. Models that only saw the original collocations before will naturally suffer from Out-of-Vocabulary (OOV) problems.

In this paper, we propose a lightweight method, Multi-feature Fusion Embedding for Chinese Named Entity Recognition (MFE-NER), which fuses extra glyph and phonetic features to detect possible substi-

©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

* This work was partly done at Shanghai Jiao Tong University and The University of Melbourne.
† Corresponding Author

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1112-1122, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

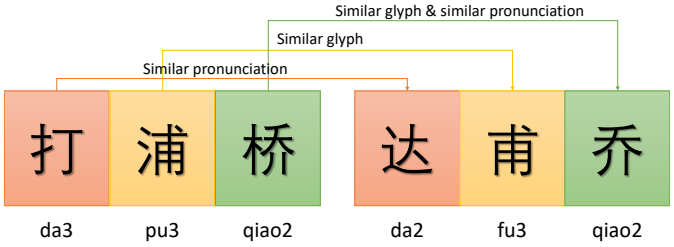



Figure 1: The substitution example of a Chinese word. On the left is a famous place in Shanghai. On the right is a new word after character substitution, which is more like a person or a brand.

tution forms of named entities in Chinese. On top of using pre-trained models to represent the semantic feature, we choose a structure-based encoding method, known as 'Five-Strokes', to introduce glyph features. In the phonetic domain, we propose 'Trans-pinyin', which combines 'Pinyin', a unique phonetic system, with international standard phonetic symbols. Experiments on three general datasets and our special-designed dataset with character substitutions have illustrated that with limited extra cost, MFE-NER can not only help NER models handle the character substitution problem but also enhance the overall performance of NER models, which makes it especially suitable to be used in current language environments.

To summarize, our major contributions are:

- We emphasize the influence of character substitution in the NER task and propose to introduce extra information from Chinese characters to handle this problem.
- For glyph features of Chinese characters, we apply the Five-Stroke encoding method, denoting structure patterns of Chinese characters.
- To represent phonetic features of Chinese characters, we propose a new method named 'Trans-pinyin', to make it possible to evaluate phonetic similarity among Chinese characters.
- Experiments show that within limited extra cost, our method is more reliable in recognizing substituted Chinese Named Entities while also slightly improving the overall performance of NER models.

## 2 Related Work

After the stage of statistical machine learning algorithms, Named Entity Recognition has stepped into the era of deep neural networks. Researchers started to use Recurrent Neural Network (RNN) (Hammerton, 2003) to recognize named entities in sentences based on character embedding and word embedding like word2vec (Church, 2017), solving the feature engineering problems that traditional statistical methods have. However, traditional RNNs may have the problem of gradient explosion and can not capture the long dependency well. In this case, the Bidirectional Long Short Term Memory (Bi-LSTM) network (Huang et al., 2015) was then applied in Chinese Named Entity Recognition, which became one of the baseline models in the Chinese NER task.

These years, large-scale pre-trained language models based on Transformer (Vaswani et al., 2017) have shown their superiority in Natural Language Processing tasks. The self-attention mechanism can better capture the long-distance dependency in sentences, and the parallel design is suitable for mass computing. Bidirectional Encoder Representations from Transformers (BERT) (Kenton and Toutanova, 2019) has achieved great success in many branches of NLP. In the Chinese Named Entity Recognition field, these pre-trained models have greatly improved the recognition performance (Cai, 2019).

CCL 2024

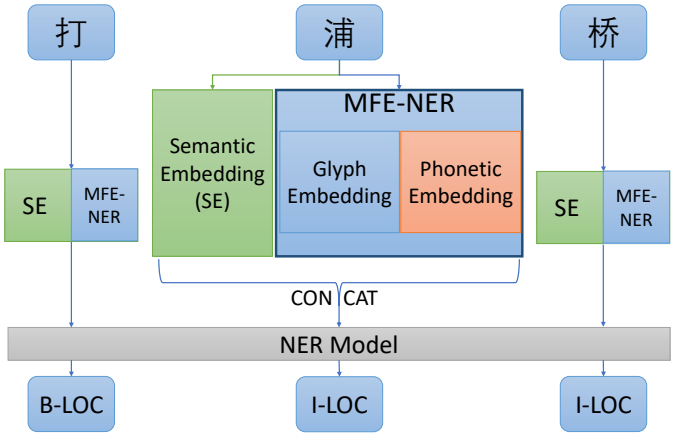



Figure 2: The structure of MFE-NER.

However, the situation is more complicated in real-language environments. Pre-trained language models do not guarantee the robustness of NER models. Researchers have started to introduce prior knowledge to improve the generalization of NER models. SMedBERT (Zhang et al., 2021) introduces knowledge graphs to help the model acquire the medical vocabulary explicitly. Meanwhile, in order to relieve character substitution problems and enhance the robustness of NER models, researchers have also paid attention to utilizing glyph and phonetic features of Chinese characters. Jiang Yang and Hongman Wang suggested using the 'Four-corner' code, a radical-based encoding method for Chinese characters, to represent glyph features of Chinese characters (Yang et al., 2021), showing the advantage of introducing glyph features in Named Entity Recognition. However, the 'Four-corner' code is not that expressive because it only works when Chinese characters have the same radicals. Tzu-Ray Su and Hung-Yi Lee suggested using Convolution Auto Encoder to build a bidirectional injection between images of Chinese characters and pattern vectors (Su and Lee, 2017). This method is brilliant but lacks certain supervision, which means that it's hard to explain the concrete meanings of the pattern vectors. Fan Yang, Jianhu Zhang et al introduced 'Five-Stroke' in NER models, which clearly improves the performance (Yang et al., 2018). Recently, Zijun Sun, Xiaoya Li, et al. proposed ChineseBERT (Sun et al., 2021), fusing glyph and phonetic features to the pre-trained models. Their work is impressive in fusing glyph and phonetic features as input for pre-trained models, but problems still exist. For example, using flattening character images is inefficient for adding huge costs to feature dimensions and enlarging possible negative influence on the model. As Chinese characters have a limited number of components in structures, it will be more convenient and efficient to extract and denote their glyph patterns by certain component encoding methods.

## 3 Method

Our MFE-NER is a lightweight Named Entity Recognition method fusing the glyph and phonetic feature embeddings for Chinese character substitution, which is complementary to pre-trained language models in the representation of Chinese characters. As shown in Figure 2, MFE-NER introduces an extra module, fusing glyph embedding with 'Five-Strokes' and phonetic embedding by 'Tran-Pinyin'.

Basically, for a Chinese sentence $S$ with length $n$, the sentence $S$ is divided naturally to different Chinese characters $S = c_1, c_2, c_3, ..., c_n$. Each character $c_i$ will be mapped to an embedding vector $\mathbf{e}_i$, which can be divided into the above three parts, $\mathbf{e}_i^s$, $\mathbf{e}_i^g$ and $\mathbf{e}_i^p$. Plus semantic embedding from pre-trained language models, all three embedding parts are designed based on a simple principle: similarity. In this paper, the character similarity between two Chinese characters $c_i$ and $c_j$ is defined by computing their L2 distance in the three aspects. Here, we use $s_{ij}^s$ to denote semantic similarity, $s_{ij}^g$ for glyph similarity and $s_{ij}^p$ for phonetic similarity. So, we have:

$$s_{ij}^s = \|\mathbf{e}_i^s - \mathbf{e}_j^s\| \tag{1}$$

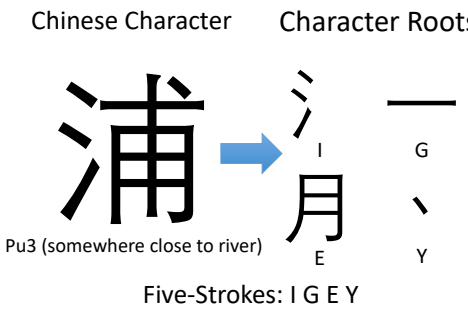



Figure 3: The 'Five-Strokes' representation of the Chinese characters, 'pu3' (somewhere close to a river). 'Five-Strokes' divides the character into four character roots ordered by writing custom so that the structure similarity can be denoted.

$$s_{ij}^{g} = \|\mathbf{e}_i^{g} - \mathbf{e}_j^{g}\| \tag{2}$$

$$s_{ij}^{p} = \|\mathbf{e}_i^{p} - \mathbf{e}_j^{p}\| \tag{3}$$

In this case, MFE-NER can better represent the inner distribution of Chinese characters in embedding space.

### 3.1 Semantic Embedding using Pre-trained Models

Semantic embedding is vital to Named Entity Recognition. In Chinese sentences, a single character does not mean a word because there is no natural segmentation in Chinese grammar. So, technically, we have two choices in acquiring Chinese embedding. The first way is word embedding, which tries to separate Chinese sentences into words and get the vector representation of words. This makes sense but is limited by the accuracy of word segmentation tools. The other is character embedding, which maps Chinese characters to different embedding vectors in semantic space. In practice, it performs better in Named Entity Recognition. Our work also requires the use of Chinese character embeddings because glyph and phonetic embeddings are targeted at single Chinese characters.

So far, pre-trained models have been widely used in the semantic domain. One typical method is Word2Vec (Mikolov et al., 2013), which starts to use static embedding vectors to represent Chinese characters in the semantic domain. Now, we have more options. BERT (Kenton and Toutanova, 2019) has its Chinese version and can express semantic features of Chinese characters more accurately, hence having better performance in NER tasks.

### 3.2 Glyph Embedding with Five-Strokes

Chinese characters, different from Latin Characters, are pictographs, which show their meanings in shapes. However, it is extremely hard for people to encode these Chinese characters in computers. A common strategy is to give every Chinese character a unique hexadecimal string, such as 'UTF-8' and 'GBK'. However, this kind of strategy processes Chinese characters as independent symbols, totally ignoring the structural similarity among Chinese characters. In other words, the closeness in the hexadecimal string value can not represent the similarity in their shapes. Some work has tried to use images of Chinese characters as glyph embedding, which is also unacceptable and ineffective due to the complexity and the large space it will take.

In this paper, we propose to use 'Five-strokes', a famous structure-based encoding method for Chinese characters, to get our glyph embedding. 'Five-Strokes' was put forward by Yongmin Wang in 1983. This special encoding method for Chinese characters is based on their structures. 'Five-Strokes' holds the opinion that Chinese characters are made of five basic strokes: horizontal stroke, vertical stroke, left-falling stroke, right-falling stroke and turning stroke. Based on that, it gradually forms a set of character roots, which can be combined to make up the structure of any Chinese character. After simplification for typing, 'Five-Strokes' maps these character roots into 25 English characters ('z' is left out) and each Chinese character is made of at most four corresponding English characters, which makes it easy to

CCL 2024

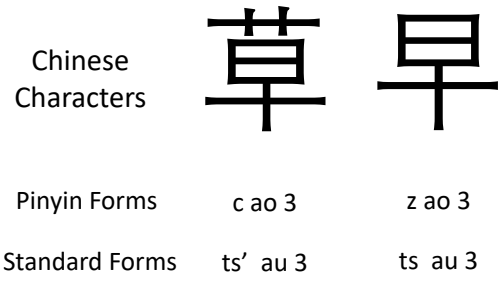



Figure 4: The 'Pinyin' form and standard form of two Chinese characters, 'cao3' (grass) on the left and 'zao3' (early) on the right.

acquire and type in computers. It is really expressive that four English characters can have $25^4 = 390625$ arrangements, while we only have about 20 thousand Chinese characters. In other words, 'Five-Strokes' has a rather low coincident code rate for Chinese characters.

For example, in Figure 3, the Chinese character 'pu3' (somewhere close to a river) is divided into four different character roots by Chinese character writing custom, which will later be mapped to English characters so that we can further encode them by introducing one-hot encoding. For each character root, we can get a 25-dimension vector. In this paper, in order to reduce space complexity, we sum up these 25-dimension vectors as the glyph embedding vector. We also list two different characters, 'fu3' (an official position) and 'qiao2' (bridge), and calculate the similarity between them. The two characters 'pu3' and 'fu3' with similar components are close in embedding space, while 'qiao2' and 'pu3' are much more distant, which gives NER models extra patterns.

### 3.3 Phonetic Embedding with Trans-pinyin

Chinese use 'Pinyin', a special phonetic system, to represent the pronunciation of Chinese characters. In the phonetic system of 'Pinyin', we have four tunes, six single vowels, several plural vowels, and auxiliaries. Every Chinese character has its expression, also known as a syllable, in the 'Pinyin' system. A complete syllable is usually made of an auxiliary, a vowel, and a tune. Typically, vowels appear on the right side of a syllable and can exist without auxiliaries, while auxiliaries appear on the left side and must exist with vowels.

However, the 'Pinyin' system has an important defect. Some similar pronunciations are denoted by totally different phonetic symbols. For the example in Figure 4, the pronunciations of 'cao3' (grass) and 'zao3' (early) are quite similar because the two auxiliaries 'c' and 'z' sound almost the same that many native speakers may confuse them. This kind of similarity can not be represented by phonetic symbols in the 'Pinyin' system, where 'c', and 'z' are independent auxiliaries. In this situation, we have to develop a method to combine 'Pinyin' with another standard phonetic system, which can better describe characters' phonetic similarities. Here, the international phonetic system seems the best choice, where different symbols have relatively different pronunciations so that people will not confuse them.

We propose the 'Trans-pinyin' system to represent character pronunciation, in which auxiliaries and vowels are transformed to standard forms and keep the tune in the 'Pinyin' system. After transformation, 'c' becomes '$ts$' and 'z' becomes '$ts'$', which only differ in phonetic weight. We also made some adjustments to the existing mapping rules so that similar phonetic symbols in 'Pinyin' could have similar pronunciations. By combining 'Pinyin' and the international standard phonetic system, the similarity among Chinese characters' pronunciations can be well described and evaluated.

In practice, we use the 'pypinyin' [1] library to acquire the 'Pinyin' form of a Chinese character. Here, we will process the auxiliary, vowel, and tune separately and normally concatenate them together after being processed.

- For auxiliaries, they will be mapped to standard forms, which have at most two English characters

[1] https://github.com/mozillazg/python-pinyin

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1112-1122, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

and a phonetic weight. We apply one-hot encoding to them so that we get two one-hot vectors and a one-dimension phonetic weight. Then we add up the two English characters' one-hot vectors, and the phonetic weight here will be concatenated to the tail.

- For vowels, they are also mapped to standard forms. However, it is a little different here. We have two different kinds of plural vowels. One is purely made up of single vowels, such as '$au$', '$eu$' and '$ai$'. The other kind is like '$an$', '$a\eta$' and '$i\eta$', which are combinations of a single vowel and a nasal voice. Here, single vowels are encoded to 6-dimension one-hot vectors and nasal voices to 2-dimension one-hot vectors, respectively. Finally, We concatenate them together, and if the vowel does not have a nasal voice, the last two dimensions will all be zero.

- For tunes, they can be simply mapped to four-dimension one-hot vectors.

## 4 Experiments

We conduct experiments on our substitution dataset and three general datasets to verify our method from different perspectives. Standard precision, recall and F1 scores are calculated as evaluation metrics to show the performance of different model settings. In this paper, we set up experiments on Pytorch and FastNLP [2] structure. All experiments are run on Nvidia RTX 3090 and 3080.

### 4.1 Dataset

| Dataset | Sentences | | |
|---|---|---|---|
| | Train | Test | Dev |
| Resume | 3821 | 463 | 477 |
| Ontonote | 15724 | 4346 | 4301 |
| Weibo | 1350 | 270 | 270 |
| Substitution | 14079 | 824 | 877 |

Table 1: Dataset Composition.

In order to verify whether our method has the ability to cope with the character substitution problem, we also built our own dataset. This specially designed dataset is collected from informal news reports and blogs. Specifically, We construct the Substitution dataset by manually labelling and replacing entities with known substitutions. Then, we randomly change the characters in the entities with similar characters. Meanwhile, the substitution dataset itself contains some sensitive materials that might violate media censorship. Generally, we ensure that every entity in this dataset is processed and replaced by similar characters. In this case, the dataset is made of pairs of original entities and their character substitution forms. This dataset consists of 15780 sentences in total and is going to test our method in an extreme language environment.

We also conduct experiments on three general NER datasets, Chinese Resume (Zhang and Yang, 2018), Ontonote (Weischedel et al., 2011), and Weibo (Peng and Dredze, 2015; He and Sun, 2017). When introducing new features, it is important to make sure that the overall performance is not negatively affected.

Among three general datasets, Chinese Resume is mainly collected from resume materials. The named entities in it are mostly people's names, positions, and company names, which all have strong logic patterns. Ontonote mainly selects corpus from official news reports, whose grammar is formal and vocabulary is quite common. Different from the above three datasets, Weibo is from social media, which includes many informal expressions. Details of the above four datasets are described in Table 1.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1112-1122, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

| Items | Range |
|---|---|
| batch size | 64 |
| epochs | 60 |
| lr | 2e-3 |
| optimizer | Adam |
| dropout | 0.4 |
| Early Stop | 5 |
| lstm layer | 1 |

Table 2: Hyper-parameters

### 4.2 Experiment Settings

The backbone of the NER model used in our work is mainly BiLSTM + CRF. The BiLSTM+CRF model is stable and has been verified in many research projects. Meanwhile, as our method is focused on providing a complementary lightweight module for current named entity recognition models, we select two pre-trained language models, static embedding trained by Word2Vec(Mikolov et al., 2013) and BERT(Kenton and Toutanova, 2019). To be more specific, BERT-base is used in our work, which has 12 transformer layers in total. Our MFE-NER with glyph and phonetic embedding is attached to them by simple concatenation.

Table 2 lists some of the hyper-parameters in our training stage. Adam is used as the optimizer, and the learning rate is set at 0.002. In order to reduce over-fitting, we set a rather high dropout (Srivastava et al., 2014) rate of 0.4. Meanwhile, the Early Stop is also deployed, allowing 5 epochs of loss not decreasing.

## 5 Results and Discussions

In this section, we first show why our method is lightweight and explain the improvement that MFE-NER achieved in recognizing substitution forms of named entities. Then, we will analyse the overall performance enhancement by applying MFE-NER. Here, the embeddings without glyph and phonetic features are named 'pure' embeddings. Basically, we compare pure embeddings from pre-trained language models with those using MFE-NER to enhance the performance of the NER task. To control variables, we also introduce the ablation study to verify the separate contribution from glyph and phonetic features. Table 3 and 4 show all the results of different model settings in different datasets.

### 5.1 A Lightweight Method

As a lightweight method, MFE-NER provides a convenient solution for all pre-trained models within limited extra cost. Different from other models utilizing complex techniques for data fusion, we simply alter the input layer, enlarging the dimension of character embedding by concatenation. Compared to pure semantic embeddings, MFE-NER only increases less than 5 % of the training time and 0.4 % of the model parameters. It is remarkable that with less than 50 dimensions added to the character embedding in the input layer, MFE-NER can cope with substitution forms while slightly improving the performance of pre-trained language models in all four datasets, showing competitive F1 scores.

### 5.2 Character Substitution Detection

Considering our substitution dataset, MFE-NER brings remarkable advances. For different semantic embeddings, MFE-NER achieves much better performance than pure embedding, improving at least 1.0 F1 score.

For the case study, there is an interesting example shown in Figure 5. The two sentences are drawn from our dataset with the same meaning, 'he wants to go to Dapuqiao'. Here, 'Dapuqiao' is a famous place in Shanghai. However, the sentence below is different because the characters in the original entity have been changed, but the word still refers to the same. Considering the model using pure semantic

[2]https://github.com/fastnlp/fastNLP




Figure 5: An example is drawn from our dataset, 'he wants to go to Dapuqiao'. The sentence at the top of the figure is the original sentence, while the sentence at the bottom is after character substitution. The model using MFE-NER gives the correct prediction.

| Models | Weibo | | | Substitution | | |
|---|---|---|---|---|---|---|
| | Precision | Recall | F1 | Precision | Recall | F1 |
| pure Word2Vec | 67.19 | 40.67 | 50.67 | **84.60** | 71.24 | 77.35 |
| glyph Word2Vec | 62.54 | 43.54 | 51.34 | 80.80 | 74.66 | 77.61 |
| phonetic Word2Vec | 66.06 | 43.30 | 52.31 | 84.09 | 74.81 | 79.18 |
| MFE-NER (Word2Vec) | **67.51** | **44.74** | **53.81** | 84.36 | **75.56** | **79.72** |
| pure BERT | 68.48 | 63.40 | 65.84 | 88.67 | 81.67 | 85.03 |
| glyph BERT | 71.88 | 64.83 | 68.18 | 89.39 | **82.86** | 86.01 |
| phonetic BERT | 70.30 | **66.27** | **68.22** | **90.48** | 82.12 | 86.09 |
| MFE-NER (BERT) | **73.54** | 63.16 | 67.95 | 90.08 | 82.56 | **86.16** |

Table 3: Results on two datasets in informal language environments, Weibo and Substitution.

embedding, it fails to give the perfect prediction for the sentence below, while the model using MFE-NER can exploit the extra information, thus giving the perfect prediction.

MFE-NER also shows its advantages in some informal language environments. The Weibo dataset is a typical example of social media. On the Weibo dataset, MFE-NER achieves 53.81 and 67.95 in models with static embedding and BERT, respectively, significantly enhancing the performance of pre-trained language models.

### 5.3 Overall NER Performance

Experiments on the other two general datasets from formal language environments also show that MFE-NER brings slight improvement to the overall NER task. On the Chinese Resume dataset, MFE-NER with BERT achieves a 95.73 F1 score, and MFE-NER with static embedding gets a 94.23 F1 score, improving the performance of pure semantic embedding from pre-trained language models by about 0.5 with respect to F1 score. On Ontonote, MFE-NER also increases the F1 score of static embedding from 80.73 to 81.24 and boosts the performance of using BERT as semantic embedding. However, owing to fewer grammar mistakes and substitution forms in the standard language environment, the performance of MFE-NER is limited.

Slight improvement still makes sense. Considering that, in Chinese, named entities have their own glyph and phonetic patterns. For example, in the glyph domain, characters in Chinese names usually contain a special character root, which denotes 'people'. What's more, characters representing places and objects also include certain character roots, which show materials like water, wood, and soil. These character roots can be utilized in 'Five-Strokes', thus improving the overall NER performance.

Based on the experiment results above, MFE-NER is able to reduce the negative influence of the character substitution phenomenon in Chinese Named Entity Recognition while slightly improving the overall performance of NER models. It makes sense that MFE-NER is suitable for solving character

| Models | Chinese Resume | | | Ontonotes | | |
|---|---|---|---|---|---|---|
| | Precision | Recall | F1 | Precision | Recall | F1 |
| pure Word2Vec | 93.97 | 93.62 | 93.79 | 72.92 | 57.83 | 64.51 |
| glyph Word2Vec | 93.87 | 93.93 | 93.90 | 71.34 | 59.88 | **65.11** |
| phonetic Word2Vec | 93.58 | 93.93 | 93.75 | **73.39** | 57.12 | 64.24 |
| MFE-NER (Word2Vec) | **94.29** | **94.17** | **94.23** | 70.92 | **59.88** | 64.93 |
| pure BERT | 94.60 | 95.64 | 95.12 | 80.53 | 80.93 | 80.73 |
| glyph BERT | 94.83 | 95.58 | 95.20 | **83.44** | 79.49 | **81.42** |
| phonetic BERT | 95.47 | **95.77** | 95.62 | 81.30 | 79.85 | 80.57 |
| MFE-NER (BERT) | **95.76** | 95.71 | **95.73** | 80.61 | **81.87** | 81.24 |

Table 4: Results on the datasets in formal language environments, Chinese Resume and Ontonotes.

substitution problems because glyph and phonetic features are introduced, which provide more prior knowledge. These extra features are complementary to semantic embedding from pre-trained models and bring information that offers more concrete proofs to NER models.

### 5.4 Ablation Study

Ablation study is thus conducted to investigate how glyph and phonetic features bring improvement by themselves. We separately add glyph embedding or phonetic embedding to pre-trained language models for comparisons. From the results on all four datasets, it is clear that models with the glyph or phonetic embedding almost all perform better than models with pure semantic embedding, which means that extra patterns from glyph and phonetic domains are all helpful in the NER task, strengthening the original model ability.

However, in some datasets, the models adding glyph or phonetic embedding only to BERT achieve higher F1 scores than MFE-NER. The reasons are to follow. On one hand, named entities in different datasets may rely on one of our provided features much more than the other features. So, in the test stage, this specific feature would contribute more to identifying named entities in the corpus. On the other hand, when we add extra information to the existing pre-trained language models, considering the different data distributions of train, dev, and test set, the model we select might give higher weights to the learned patterns from glyph or phonetic domain, which shows good performance on dev set but possibly leads to over-fitting.

## 6 Conclusion

Nowadays, the informal language environment created by social media has deeply changed the way that people express their thoughts. Using character substitution to generate new named entities to avoid social media filtering has become a common linguistic phenomenon that is a big challenge for NER. In this paper, we propose a lightweight method fusing the glyph and phonetic features for Chinese Named Entity Recognition (MFE-NER) to handle the Chinese character substitution problem, which introduces 'Five-strokes' to represent glyph features, making it more convenient to differentiate similar characters, and develops a system named 'Trans-Pinyin' to represent phonetic features of Chinese characters, solving the problem of similar pronunciations in Pinyin. Experiments show that our model, MFE-NER, can assist pre-trained models and reduce the influence caused by character substitution.

## Acknowledgments

We thank anonymous reviewers for their helpful comments. This project is supported by the National Natural Science Foundation of China (No. 61772337). Thanks, Bin He from Tencent, for sharing his opinions on the experiment settings.

## References

Qing Cai. 2019. Research on chinese naming recognition model based on bert embedding. In *2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS)*, pages 1–4. IEEE.

Kenneth Ward Church. 2017. Word2vec. *Natural Language Engineering*, 23(1):155–162.

Lea Frermann, Jiatong Li, Shima Khanehzar, and Gosia Mikolajczak. 2023. Conflicts, villains, resolutions: Towards models of narrative media framing. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8712–8732.

James Hammerton. 2003. Named entity recognition with long short-term memory. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*, pages 172–175.

Hangfeng He and Xu Sun. 2017. F-score driven max margin neural network for named entity recognition in chinese social media. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 713–718.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv e-prints*, pages arXiv–1508.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Xiaonan Li, Hang Yan, Xipeng Qiu, and Xuanjing Huang. 2020. Flat: Chinese ner using flat-lattice transformer. *arXiv preprint arXiv:2004.11795*.

Jiatong Li, Wei Liu, Zhihao Ding, Wenqi Fan, Yuqiang Li, and Qing Li. 2024a. Large language models are in-context molecule learners. *arXiv preprint arXiv:2403.04197*.

Jiatong Li, Yunqing Liu, Wenqi Fan, Xiao-Yong Wei, Hui Liu, Jiliang Tang, and Qing Li. 2024b. Empowering molecule discovery for molecule-caption translation with large language models: A chatgpt perspective. *IEEE Transactions on Knowledge and Data Engineering*.

Qinghua Mao, Jiatong Li, and Kui Meng. 2022. Improving chinese named entity recognition by search engine augmentation. *arXiv preprint arXiv:2210.12662*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Computer Science*.

Nanyun Peng and Mark Dredze. 2015. Named entity recognition for chinese social media with jointly trained embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 548–554.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Tzu-Ray Su and Hung-Yi Lee. 2017. Learning chinese word representations from glyphs of characters. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 264–273.

Zijun Sun, Xiaoya Li, Xiaofei Sun, Yuxian Meng, Xiang Ao, Qing He, Fei Wu, and Jiwei Li. 2021. Chinesebert: Chinese pretraining enhanced by glyph and pinyin information. *arXiv e-prints*, pages arXiv–2106.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Martha Palmer, Nianwen Xue, Mitchell Marcus, Ann Taylor, Craig Greenberg, Eduard Hovy, Robert Belvin, et al. 2011. Ontonotes release 4.0. *LDC2011T03, Philadelphia, Penn.: Linguistic Data Consortium*.

Fan Yang, Jianhu Zhang, Gongshen Liu, Jie Zhou, Cheng Zhou, and Huanrong Sun. 2018. Five-stroke based cnn-birnn-crf network for chinese named entity recognition. In Min Zhang, Vincent Ng, Dongyan Zhao, Sujian Li, and Hongying Zan, editors, *Natural Language Processing and Chinese Computing*, pages 184–195, Cham. Springer International Publishing.

CCL 2024

Jiang Yang, Hongman Wang, Yuting Tang, and Fangchun Yang. 2021. Incorporating lexicon and character glyph and morphological features into bilstm-crf for chinese medical ner. In *2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE)*, pages 12–17. IEEE.

Yue Zhang and Jie Yang. 2018. Chinese ner using lattice lstm. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1554–1564.

T. Zhang, Z. Cai, C. Wang, M. Qiu, and X. He. 2021. Smedbert: A knowledge-enhanced pre-trained language model with structured semantics for medical text mining. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

CCL 2024

# UDAA: An Unsupervised Domain Adaptation Adversarial Learning Framework for Zero-Resource Cross-Domain Named Entity Recognition

**Baofeng Li[2], Jianguo Tang[1,†], Yu Qin[2], Yuelou Xu[2], Yan Lu[2], Kai Wang[1] Lei Li[1,*], Yanquan Zhou[1,*]**
[1]Beijing University of Posts and Telecommunications
[2]China Electric Power Research Institute Company Limited
{jianguotang, wk, leili} @bupt.edu.cn
libaofeng@epri.sgcc.com.cn

## Abstract

The zero-resource cross-domain named entity recognition (NER) task aims to perform NER in a specific domain where labeled data is unavailable. Existing methods primarily focus on transferring NER knowledge from high-resource to zero-resource domains. However, the challenge lies in effectively transferring NER knowledge between domains due to the inherent differences in entity structures across domains. To tackle this challenge, we propose an Unsupervised Domain Adaptation Adversarial (**UDAA**) framework, which combines the masked language model auxiliary task with the domain adaptive adversarial network to mitigate inter-domain differences and efficiently facilitate knowledge transfer. Experimental results on CBS, Twitter, and WNUT2016 three datasets demonstrate the effectiveness of our framework. Notably, we achieved new state-of-the-art performance on the three datasets. Our code will be released.

## 1 Introduction

Named entity recognition (NER) is a fundamental task in natural language processing. However, in real-world scenarios, obtaining the amount of domain-specific labeled data for the NER task is often challenging, as it can be expensive and time-consuming (Ma and Hovy, 2016; Lample et al., 2016; Akbik et al., 2018; Winata et al., 2019). Hence, zero-resource cross-domain NER (Jia et al., 2019; Bari et al., 2020; Zhang et al., 2021; Karouzos et al., 2021; Zheng et al., 2022), which addresses the issue of data scarcity by adapting NER models to specific domains without available labeled data, draws more and more attention.

Based on the mainstream methods, recent researches have focused on the transfer of NER knowledge from high-resource to zero-resource domains. These methods typically involve training a source model on labeled data from the source domain to learn domain invariant features and apply the adapting NER models to the target domain. In the process of transferring knowledge to the target domain, the Masked Language Model(MLM) is currently recognized as an unsupervised adaptive auxiliary task with good performance. AdaptBERT (Han and Eisenstein, 2019) fine-tunes the MLM on unlabeled data to learn both task-special knowledge and domain-invariant knowledge. Furthermore, TOF (Zhang et al., 2021) addresses the zero-resource cross-domain task scenario through MLM capturing enough NER task knowledge. The acquired knowledge serves as a bridge to facilitate the transfer of NER knowledge.

MLM plays a significant role in extracting context and recognizing entity boundaries. However, during knowledge transfer, especially when mapping between two domains, the transfer of knowledge from different tasks and the varying data distributions across domains can introduce noise and hinder model generalization performance. For instance: a) I'm watching a documentary about a **shark** on TV, and b) The stock market experienced a **shark** attack as prices plummeted. In the social media domain, the word **shark** represents sea animals. However, in the news domain, this is used metaphorically, referring to the stock market crash. The word "shark" shares the same affixes and root knowledge across different domains, yet it carries different semantic meanings in each domain. Meanwhile, due to such

[†]Indicates co-first author.
[*]Corresponding authors.

©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1123-1135, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

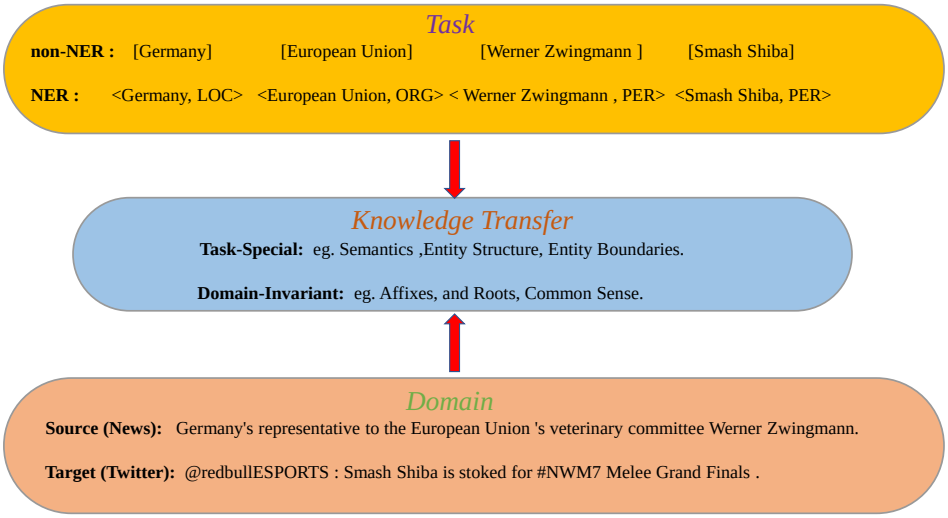



Figure 1: The example of knowledge transfer process from two perspectives: domain and task.

semantic biases across different domains, the MLM task may fail to adapt to the specific context and task requirements of the target domain, resulting in a decrease in model performance in the target domain.

Inspired by the idea of (Bari et al., 2020), which adopted the adversarial approach to learning language-invariant knowledge in an unsupervised adaptive way, we propose a framework named Unsupervised Domain Adaptation Adversarial (UDAA). UDAA combines the MLM auxiliary task with the Domain Adaptation Adversarial Network (DAAN) to mitigate inter-domain disparities. By incorporating an adversarial loss, the model becomes more adept at distinguishing domain-invariant knowledge from task-specific knowledge. This encourages the development of resource-agnostic representations, allowing knowledge from high-resource domains to be more effectively aligned with zero-resource domains.

As shown in Figure 1, in our work, the knowledge transfer process is from two perspectives: domain and task.

**Domain.** From the domain perspective, the domain-invariant knowledge can be transferred, such as affixes, roots, and common sense, in both source and target domains. To ensure the consistency between feature representations extracted from the source domain and those from the target domain, we employ the DAAN method to align features across different domains, eliminating distribution discrepancies between domains. This approach captures domain-relevant knowledge, allowing for the learning of more domain-invariant features to ensure the model's generalization performance.

**Task.** From the task perspective, the task-special knowledge can be transferred, such as entity structure and senmantics, in both NER and non-NER task. These task-specific features play a crucial role in the complementary process of NER and non-NER task(Rei, 2017; Peters et al., 2017). For instance, by introducing the MLM task, the model can extract entity and semantic information from the context and identify entity boundaries, thus compensating for missing entity information in domain-specific NER task. In the UDAA model, tasks in different stages act as bridges from high-resource domains to zero-resource domains, facilitating effective knowledge transfer.

In this paper, the UDAA framework is designed to address the challenge of zero-resource cross-domain NER task, which focuses on ensuring the adaptability of unsupervised domains by combining the MLM auxiliary task with DAAN. This integration enables the framework to transfer knowledge from both domain and task perspectives. With enough training on the labeled data from the source domain, the framework generates pseudo-data on the unlabeled corpus to address the data scarcity challenge encountered in the target domain. The training process of the UDAA framework consists of three unified steps: Adversarial Domain Learning (ADL), Adaptive Pseudo-Data Generation(APG), and Domain Adaptive

CCL 2024

Prediction(DAP).

To validate the effectiveness and superiority of our proposed approach, we conduct experiments on three datasets for zero-resource cross-domain NER task: CBS (Jia et al., 2019), Twitter (Zhang et al., 2018b), and WNUT2016 (Strauss et al., 2016).

**Our contributions can be summarized as follows:**

- We propose the UDAA framework, which performs the zero-resource cross-domain NER task by unifying training through the steps of Adversarial Domain Learning, Adaptive Pseudo-Data Generation, and Domain Adaptive Prediction.
- We combine the MLM auxiliary task with the DAAN to mitigate inter-domain disparities. By incorporating an adversarial loss, the model becomes more adept at distinguishing task-special knowledge from domain-invariant knowledge. This integration provides a lightweight structure for the UDAA framework.
- Experimental results on the CBS, Twitter, and WNUT2016 datasets demonstrate the effectiveness of our UDAA framework. Meanwhile, we achieve new state-of-the-art performance on the three datasets.

## 2 Related Work

**Cross-domain NER.** Some studies (Devlin et al., 2018; Sun et al., 2020; Liu et al., 2021) achieves remarkable performance in cross-domain NER scenario, yet they require many domain-related labeled data for training when adapting to the target domain. To this end, cross-domain NER algorithms (Yang et al., 2018; Lin and Lu, 2018; Liu et al., 2020b) that alleviate the data scarcity issue and boost the models' generalization ability to target domain have drawn substantial attention recently. To solve this problem, some studies introduce multitask learning (Liu et al., 2020b; Wang et al., 2020) or designing new model architectures (Jia et al., 2019; Liu et al., 2020c; Jia and Zhang, 2020) for improving the NER performance of the target domain by training on data from both source and target domain. However, these methods typically require training with vast amounts of labeled source domain data to achieve satisfactory performance in the target domain. Our work differs in that we neither requires a large amount of domain-related data nor the design of a new model structure. We design a unified training framework and use models of different tasks as the bridge from high-resource to zero-resource domain to achieve knowledge transfer.

**Zero-resource NER.** Some studies (Jia and Zhang, 2020; Pfeiffer et al., 2020) focus on enhancing the architectural design of existing models by incorporating new components to capture specific knowledge, such as entity types and task characteristics. Different from these methods, our approach only modifies the training procedure without changing model structures. Other studies introduce different auxiliary tasks to alleviate data scarcity (Han and Eisenstein, 2019; Phang et al., 2020; Xue et al., 2020a). They are usually based on multi-task learning. Multi-task learning requires balance between the NER task and auxiliary tasks, which needs carefully designed objectives. In addition, LLMs (Ouyang et al., 2022; Zeng et al., 2022a) have good learning and expression capabilities, so they are also used in zero-resource NER task and have achieved some good results. Notably, our work differs in that we design an unsupervised domain adaptation fine-tuning framework to exploit more diverse data and training strategies.

**Adversarial Learning.** Adversarial learning originates from Generative Adversarial Nets(GAN) (Goodfellow et al., 2014). Adversarial learning is a regularization method for improving the generalization of a classifier. It does so by improving model's robustness to adversarial examples, which are created by making small perturbations to the input. Recently, many studies (Bari et al., 2020; Zhao et al., 2022) have tried to apply adversarial learning to NLP tasks. They(Liu et al., 2017) extended adversarial training to the multi-task learning framework for text classification, aiming to alleviate the domain-invariant (shared) and task-special (private) latent feature spaces from interfering with each other. Notably, Our method integrates both the adversarial training in an unified framework to find domain-invariant information for MLM tasks.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1123-1135, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

## 3 Our Approach

In recent years, researchers have paid more attention to cross-language scenarios, as the data in cross-domain scenarios is more difficult to obtain. Therefore, we adopt TOF framework (Zhang et al., 2021) as our baseline, which was designed for unsupervised domain adaptation in sequence labeling tasks and achieved state-of-the-art performance in previous work.

The UDAA framework is enlightened by the TOF. Our approach consists of the knowledge transfer process and the fine-tuning process. The architecture of UDAA is illustrated in Figure 2.

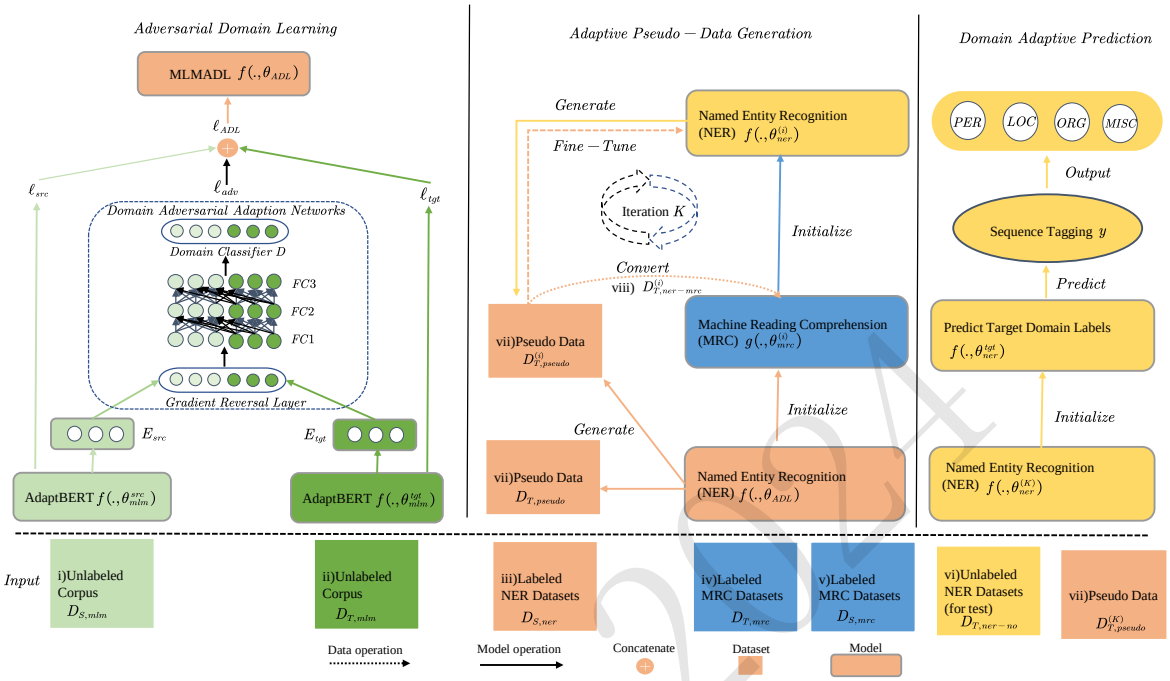



Figure 2: The overall architecture of the UDAA framework. The dotted arrows represent the knowledge transfer process of the data, and the solid arrows represent the fine-tuning process of the model. At the ADL stage, we combine the AdaptBERT model with the DAAN and incorporating the adversarial loss to distinguish domain-invariant knowledge from task-specific knowledge. At this time, an MLMADL model will be generated as the input of the next APG stage. In the APG stage, the MRC framework (Li et al., 2019) will be used for $K$ rounds of iterations to generate high-quality pseudo-data and the best NER model for prediction in the next DAP stage. The models for tasks at different stages serve as bridges from high-resource domains to zero-resource domains and enable knowledge transfer.

### 3.1 Knowledge Transfer Process

As shown in Figure 2, our framework aims to facilitate knowledge transfer to the target domain, which displays how to transfer not only domain but also task knowledge from various data. Relevant information about the data is listed in Table 1.

We consider six kinds of datasets: i) Unlabeled Corpus $D_{S,mlm}$, ii) Unlabeled Corpus $D_{T,mlm}$, iii) Labeled NER Dataset $D_{S,ner}$, iv) Labeled MRC Dataset $D_{T,mrc}$, v) Labeled MRC Dataset $D_{S,mrc}$, and vi) Unlabeled NER Dataset (for test) $D_{T,ner-no}$, where datasets $\{$ i), iii), v)$\}$ belong to the source domain and datasets $\{$ ii), iv), vi)$\}$ belong to the target domain. Additionally, two important intermediate generated data are also included: vii) Pseudo Data $D_{T,pseudo}$, and viii) MRC-Style Data $D_{T,ner-mrc}$, where $D_{T,ner-mrc}$ is converted from $D_{T,pseudo}$ and plays an important role in the knowledge transfer of MRC intermediate task.

Considering the discrepancy between the source and target domains, it is essential to perform fine-tuning tasks on both the source and target data. In addition to the labeled data for the NER task, we also

| Number | Name | Domain | Task |
|---|---|---|---|
| i) | $D_{S,mlm}$ | Source | MLM |
| ii) | $D_{T,mlm}$ | Target | MLM |
| iii) | $D_{S,ner}$ | Source | NER |
| iv) | $D_{T,mrc}$ | Target | MRC |
| v) | $D_{S,mrc}$ | Source | MRC |
| vi) | $D_{T,ner-no}$ | Target | NER |
| vii) | $D_{T,pseudo}$ | Target | NER |
| viii) | $D_{T,ner-mrc}$ | Target | MRC |

Table 1: Various data descriptions in the knowledge transfer process.

utilize the non-NER task. For example, $D_{S,mlm}$ and $D_{T,mlm}$ are used for ADL stage, while $D_{S,mrc}$ and $D_{T,mrc}$ are used for APG stage.

### 3.2 Fine-tuning Process

There are three steps to performing the NER task in the fine-tuning process: a) Adversarial Domain Learning, b) Adaptive Pseudo-Data Generation, and c) Domain Adaptive Prediction.

#### 3.2.1 Adversarial Domain Learning

We employ two independent encoders to fine-tune the MLM auxiliary task to capture domain-invariant and task-special representations for source and target domains. Concretely, given two input token sequences $x_s = \{x_i\}_i^N$ and $x_t = \{x_j\}_j^M$, where $N, M$ represents the number of words in source and target domains. We feed it into the feature encoder to obtain contextualized word embeddings $E_{src}$ and $E_{tgt}$:

$$E_{src} = AdaptBERT(x_s) \tag{1}$$

$$E_{tgt} = AdaptBERT(x_t) \tag{2}$$

However, the MLM task may fail to adapt to the specific context and task requirements of the target domain due to semantic biases present across different domains, resulting in a decrease in model performance in the target domain. To address this issue, we integrate DAAN with the MLM task and conduct joint training to mitigate semantic biases between domains and facilitate effective knowledge transfer.

In our work, we use adversarial training to find domain invariant representations . The DAAN exhibit an architecture whose first few feature extraction layers are shared by two classifiers trained simultaneously. DAAN minimizes the domain classification loss with respect to parameters specific to the domain classifier $D$ , while maximizing it with respect to the parameters that are common to both classifiers. This minimax optimization becomes possible via the use of a gradient reversal layer (GRL). By incorporating DAAN, a more powerful model MLMADL is trained, which can capture more domain-invariant knowledge, thereby enhancing the model's generalization performance.

$$L_i(x_d) = W_i \cdot x_d + b_i \tag{3}$$

$$D(E_d) = \sigma([L_2 \cdot (L_1(E_d) \cdot W_3 + b_3)]) \tag{4}$$

where $d \in \{src, tgt\}$ and $L_i$ correspond to the fully connected layer in DAAN. $D(E_d)$ discriminates whether $E$ comes from the source or the target domains. $W_i$ and $b_i$ represent different fully connected layer parameters. The loss function $\ell_{adv}$ for DAAN is calculated as follows:

$$\ell_{adv} = \Sigma_{x_s} log(D(E_{src})) + \Sigma_{x_t} log(1 - D(E_{tgt})) \tag{5}$$

The total loss function $\ell_{ADL}$ for jointly train in the ADL stage is defined as follows:

$$\ell_{ADL} = \ell_{adv} + \ell_{src} + \ell_{tgt} \tag{6}$$

CCL 2024

where $\ell_{src}$ represents the source domain loss function and $\ell_{tgt}$ represents the target domain loss function.

By incorporating $\ell_{adv}$, the MLMADL model becomes more adept at distinguishing domain-invariant knowledge from task-specific knowledge. This encourages the development of resource-agnostic representations, allowing knowledge from high-resource domains to be more effectively aligned with zero-resource domains.

### 3.2.2 Adaptive Pseudo-Data Generation

We obtain a trained NER model by using labeled data from the source domain along with the initial parameters of the MLMADL, which is utilized on the unlabeled corpus in the target domain and generates pseudo-data. To iteratively optimize this pseudo-data, we employ the MRC auxiliary task. The MRC task offers several advantages: i) It enhances the NER model's capability to extract spans, thereby capturing semantic information associated with different entity types (Wu et al., 2019). ii) It acts as a bridge between NER and other tasks. iii) It facilitates the transfer of knowledge from diverse tasks into a unified training framework (Liu et al., 2020a).

In the iterative optimization process of the MRC task, the dataset format required differs from that of the NER task dataset. Therefore, dataset conversion is necessary. The dataset is represented using triples $(question, answer, context)$, and to fully leverage contextual information from each sentence, we treat the description of each entity as a query. Each entity in the dataset corresponds to an entity label, and each entity label is associated with a query $q_y = q_1, q_2, ..., q_m$, where $m$ represents the length of the generated query. The training objective of the MRC task at this point is to answer natural language queries given the context to obtain named entity information.

In the MRC task, the design of natural language queries plays a crucial role because it requires skillful construction of queries to encode prior knowledge about entity class labels. Otherwise, named entity information in the text cannot be effectively obtained. We adopt a template-based approach to construct natural language queries for relevant entity type labels.

We employ a span selection strategy (Li et al., 2019) by feeding the context word vectors of each label into two separate linear classification layers. Using binary classifiers to predict the start and end indices of each entity allows the model to determine the boundaries of the entity based on specific queries. The probability formula for computing the $start$ or $end$ indices of each label as entity spans is as follows:

$$p^{start} = softmax(W_{start} \cdot x) \tag{7}$$

$$p^{end} = softmax(W_{end} \cdot x) \tag{8}$$

where $W_{start}$ and $W_{end} \in R^{d_1 \times 2}$ are learnable parameters, and $d_1$ denotes the dimensions of contextualized word embedding.

After $K$ rounds of iterative optimization using the MRC task, we obtain high-quality pseudo-data, which helps alleviate the issue of data scarcity in the target domain. In addition, through the MRC intermediate task, we also obtained the NER model for prediction in the next DAP stage.

### 3.2.3 Domain Adaptive Prediction

In the DAP stage, the NER model inherits parameters from the APG stage and applies it to predict sequence labels in the target domain. During the DAP stage, given the text $w = (w_1, w_2, ...w_T)$ of the dataset $D_{T,ner-no}$, special characters $[CLS]$ and $[SEP]$ are respectively appended at the beginning and end of each sentence to obtain special representations. After encoding by the model $f(\cdot, \theta^{tgt}_{ner})$, a series of contextual feature representations $x$ are returned. These representations $x$ are then inputted into linear classification layers, maximizing the probability of each label $y_t$ using true entity labels. We opt for the cross-entropy loss function, and the probability calculation for entity prediction is as follows:

$$logp(y_t|w_{1:T}) = \beta_{y_t} \cdot x - logy \in Y \Sigma exp(\beta_{y_t} \cdot x) \tag{9}$$

where the contextualized embedding $x$ captures information from the entire sequence $w_{1:T} = (w_1, w_2, ..., w_T)$, and $\beta_{y_t}$ is a vector of weights for each tag $y \in \{PER, LOC, ORG, MISC\}$.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1123-1135, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

In the DAP stage, the UDAA model utilizes high-quality pseudo-labeled data refined through $K$ rounds of iterative optimization, along with a well-generalized NER model, to make predictions. This process enables the NER task model to effectively adapt to the target domain and achieve domain-adaptive sequence labeling predictions.

## 4 Experiments

### 4.1 Datasets

We take Conll-2003 for English in the news domain as our source domain dataset for the NER task. As for the target domain, we consider three datasets: a) CBS SciTech News(referred to as CBS) (Jia et al., 2019), from the technology news domain, b) Twitter (Zhang et al., 2018b), and c) WNUT2016 (Strauss et al., 2016) both from the social media domain. The unlabeled corpora from these datasets are readily available for the MLM task. In addition, for the iterative optimization process in the MRC task, we select NewsQA (Trischler et al., 2016) and TweetQA (Xiong et al., 2019) as our target domain dataset, while SQuAD (Rajpurkar et al., 2016) is chosen as the source domain dataset.

To ensure one-to-one correspondence between labels across different domains, we chose the "no-type" category from the WNUT2016 dataset, where entities don't have specific label types and only contain entity types in the BIO format. Additionally, the labels in the CBS and Twitter datasets cover four entity types: PER, LOC, ORG, and MISC. To ensure transferability and adaptability to the target domain, we applied a data preprocessing approach to remove entity label data from the training set in the CBS and Twitter datasets. All datasets in the NER task adhere to the Conll-2003 dataset standard, where data is divided into two columns: the first column contains word segmentation in the input sentence, and the second column contains named entity type labels corresponding to each word.

### 4.2 Hyperparameters

| Module | Task | CBS | Twitter | WNUT2016 |
|---|---|---|---|---|
| ADL | MLM | 2e-5 | 2e-5 | 2e-5 |
| ADL | NER | 5e-5 | 2e-5 | 5e-5 |
| APG | NER | 8e-5 | 1e-6 | 1e-6 |
| APG | MRC | 8e-6 | 3e-5 | 1e-6 |
| DAP | NER | 3e-5 | 5e-6 | 2e-6 |

Table 2: Learning rate for UDAA framework.

In the training procedure, we choose the Adam (Zhang et al., 2018a) optimizer and our experiment runs on the Tesla P100 server, which takes more than 14h. We conduct the hyperparameters search to determine the appropriate learning rate for the $ADL$ module. We explore a range of learning rate values: 5e-6, 8e-6, 1e-5, 2e-5, 3e-5, 5e-5, and 8e-5. We also explore a range of batch size values: 16, 32, and 64, which are limited by server computing power and we choose the batch size of 32. The selection of hyperparameters values is based on the best validation performance.

Furthermore, we fine-tune each MLM model for 3 epochs, MRC for 6 epochs, and NER for 6 epochs. Additionally, we perform one iteration of APG stage and achieve the best performance. The learning rates for the UDAA framework can be found in Table 2. Other hyperparameters are set following the approach described by TOF(Zhang et al., 2021).

### 4.3 Model

We evaluate the following models by P, R, and F1 scores:

- **ChatGPT**(Ouyang et al., 2022) select the GPT-3.5-Turbo model for zero-resource named entity prediction.
- **ChatGLM**(Zeng et al., 2022a) select the ChatGLM2-6B model for zero-resource named entity prediction.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1123-1135, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

| Model | CBS | | | Twitter | | | WNUT2016 | | |
|---|---|---|---|---|---|---|---|---|---|
| | P | R | F1 | P | R | F1 | P | R | F1 |
| ChatGPT†(Ouyang et al., 2022) | 28.23 | 25.63 | 26.87 | 17.64 | 19.92 | 18.71 | 15.85 | 18.05 | 16.57 |
| ChatGLM†(Zeng et al., 2022a) | 1.74 | 2.11 | 1.91 | 1.58 | 1.88 | 1.71 | 1.29 | 1.33 | 1.31 |
| Cross-LM BiLSTM(Jia et al., 2019) | 68.48 | 79.52 | 73.59 | - | - | - | - | - | - |
| mCell LSTM(Jia and Zhang, 2020) | - | - | 75.19 | - | - | - | - | - | - |
| COFEE-MRC(Xue et al., 2020b) | - | - | - | - | - | 54.56 | - | - | - |
| TOF(Zhang et al., 2021) | - | - | 76.41 | - | - | 67.94 | - | - | 67.86 |
| BiLSTM+CRF*(Liu et al., 2020b) | 78.10 | 59.94 | 67.82 | 65.72 | 58.89 | 62.12 | / | / | / |
| BiLSTM+CRF+DAAN* | 78.48 | 61.29 | 68.82 | 65.88 | 60.74 | 63.21 | / | / | / |
| AdaptBERT*(Han and Eisenstein, 2019) | 69.80 | 81.26 | 75.10 | 66.49 | 66.62 | 66.55 | 65.66 | 62.30 | 64.10 |
| TOF*(Zhang et al., 2021) | 71.11 | 82.58 | 76.42 | 68.20 | 67.61 | 67.90 | 70.56 | 62.74 | 66.42 |
| UDAA($ours$) | 71.76 | 82.23 | **76.64** | 68.94 | 68.67 | **68.81** | 71.17 | 65.53 | **68.24** |
| $w/o$ APG Stage | 71.39 | 81.97 | 76.32 | 68.03 | 67.70 | 67.86 | 69.01 | 65.39 | 67.15 |
| $w/o$ ADL Stage | 71.42 | 81.13 | 75.97 | 66.38 | 67.01 | 66.70 | 67.67 | 64.73 | 66.17 |
| $w/o$ APG Stage & $w/o$ ADL Stage | 68.83 | 79.28 | 73.69 | 65.85 | 65.10 | 65.48 | 61.44 | 63.32 | 62.37 |

Table 3: Results of our UDAA framework compared to previous state-of-the-art methods for the zero-resource cross-domain NER task. "†" indicates the prediction results of the LLMs. "-" indicates results not provided in prior methods. "*" signifies results re-implemented by us. "/" denotes cases where the processed WNUT2016 dataset is incompatible with the BiLSTM model and therefore results are unavailable. Bold text highlights the best experimental outcomes.

| Model | CBS | | | Twitter | | | WNUT2016 | | |
|---|---|---|---|---|---|---|---|---|---|
| | Time | Flops | Size | Time | Flops | Size | Time | Flops | Size |
| **TOF Framework** | | | | | | | | | |
| MRC Enhancing | 19.55h | 3656.9B | 312M | 13.53h | 2419.0B | 312M | 13.56h | 1397.8B | 312M |
| Pseudo Data Enhancing | 0.54h | 347.9B | 103.3M | 0.15h | 337.1B | 103.3M | 0.11h | 282.7B | 103.3M |
| Continual Learning Enhancing | 12.24h | 3086.3B | 206.6M | 12.35h | 2439.5B | 206.6M | 12.01h | 1432.9B | 206.6M |
| **Total** | 31.84h | 7091.1B | 624M | 25.50h | 4858.5B | 624M | 25.6h | 3113.4B | 624M |
| **UDAA Framework** | | | | | | | | | |
| Adversarial Domain Learning | 12.60h | 1043.7B | 103.3M | 1.17h | 872.6B | 103.3M | 1.28h | 726.1B | 103.3M |
| Adaptive Pseudo-Data Generation | 12.24h | 3086.3B | 206.6M | 12.35h | 2439.5B | 206.6M | 12.01h | 1432.9B | 206.6M |
| Domain Adaptive Prediction | 0.54h | 347.9B | 103.3M | 0.15h | 337.1B | 103.3M | 0.11h | 282.7B | 103.3M |
| **Total** | 25.33h | 4477.9B | 415.3M | 13.69h | 3649.2B | 415.3M | 13.71h | 2441.6B | 415.3M |

Table 4: Comparison of UDAA and TOF frameworks in computing resources. The MRC Enhancing, Pseudo Data Enhancing, and Continual Learning Enhancing are mentioned in TOF(Zhang et al., 2021).

- **Cross-LM BiLSTM**(Jia et al., 2019) using a language model as a cross-domain bridge, automatically generate training parameters for the LSTM model to perform zero-resource named entity prediction.
- **mCell LSTM** (Jia and Zhang, 2020) designed a multicell compositional LSTM for cross-domain NER task.
- **COFEE-MRC**(Xue et al., 2020b) designed a neural network model with MRC auxiliary task.
- **BiLSTM+CRF**(Liu et al., 2020b) designed a BiLSTM and CRF paradigm for zero-resource cross-domain NER.
- **BiLSTM+CRF+DAAN.** We built on the work(Liu et al., 2020b), incorporating the DAAN components.
- **AdaptBERT.**(Han and Eisenstein, 2019) adopted the domain-tuning and task-tuning in the cross-domain NER task.
- **TOF**(Zhang et al., 2021) performed MLM and MRC auxiliary tasks. It is the previous best state-of-the-art and we take the TOF as our baseline in the zero-resource cross-domain scenario.

CCL 2024

- **UDAA** The framework proposed in our work and details can be found in section 3.

## 5 Results and Analysis

### 5.1 Framework Performance

Table 3 presents the main results of our framework compared to previous methods in the zero-resource cross-domain NER task. We regard re-implemented results of TOF (Zhang et al., 2021) as our baseline. Our framework yields obvious improvements over the baseline (CBS: 0.22 ↑, Twitter: 0.91 ↑, and WNUT2016: 1.82 ↑ ) and achieves new state-of-the-art results on the three datasets.

The two selected LLMs, ChatGPT and ChatGLM, performed poorly. The reasons may be: (a) The instructions used by the LLMs require the model to generate more fine-grained entity tags. In the zero-resource scenario, this more complex instruction design makes it difficult for the models to ensure the accuracy of the output; (b) LLMs are prone to missing entities, mislabeling, and multiple entities; (c) The selected LLMs is limited by computing resources and is not fine-tuning for NER task.

In conclusion, all these results verify the effectiveness and transferability of our framework UDAA in the zero-resource cross-domain NER task.

### 5.2 Effectiveness of DAAN

As shown in Table 3, the integration of DAAN into the BiLSTM model leads to improved F1 scores in the experimental results on the CBS(1.0 ↑) and Twitter(1.09 ↑) datasets, which shows the effectiveness of DAAN in enhancing feature extractors. It also indicates that DAAN enables the model to acquire additional entity information and gain the better understanding of the target domain data.

In addition, our model builds upon the baseline of TOF, and we observed enhancements in the F1 score by introducing DAAN while omitting the MRC task during the fine-tuning process. Moreover, DAAN provides a lightweight and effective method, which also accelerates the model's training speed, consequently reducing training time and computational resource costs. We compare the computing resource consumption of the UDAA with TOF, measured by Time(time consumption, $h$ means hour), Flops(floating point operations per second, $B$ means billion), and Size(model parameters). The results shown in Table 4 demonstrate that UDAA consumes fewer computing resources than TOF.

| Method | CBS | Twitter | WNUT2016 |
|---|---|---|---|
| with DAAN | 75.97 | 66.70 | 66.17 |
| without DAAN | 75.66 | 66.04 | 65.14 |

Table 5: Compared the experimental results of NER task with and without the integration of DAAN.

Finally, as shown in Table 5, the integration of DAAN into the MLM task yielded better experimental results on the CBS(0.31 ↑), Twitter(0.66 ↑) and WNUT2016(1.03 ↑) datasets. This further underscores the role of the introduced adversarial loss helps the MLM task distinguish between domain-invariant features and task-specific features, thereby enhancing the model's robustness and generalization performance.

Overall, whether we consider the notable reduction in computational resources as demonstrated in Table 4 or the enhancements in the experimental results as evident in Table 3 and Table 5, it underscores the effectiveness of DAAN. The integration of this structure has proven to be beneficial for the framework.

### 5.3 Ablation Analysis

#### 5.3.1 Component Impact

To assess the influence of various components of UDAA, we examine the effects of two modules: i) ADL, and ii) APG. As shown in Table 3, when we remove the APG stage (CBS(0.32 ↓), Twitter(0.95 ↓) and WNUT2016(1.09 ↓)), implying the absence of high-quality pseudo-data and the MRC task, the model's performance decline is less pronounced compared to when the ADL stage (CBS(0.67 ↓), Twitter(2.11 ↓) and WNUT2016(2.07 ↓)) is absent. This highlights the substantial contribution of the introduced DAAN to the task model in subsequent stages, thus improving its robustness and generalization performance.

Meanwhile, the experimental impact of removing the APG stage alone or removing the ADL stage alone is far lower than removing both simultaneously, indicating the crucial role of unified joint training between the APG and ADL stages for the UDAA model. This may be attributed to the coupling effect of the model eliminating inter-domain differences in the ADL stage and high-quality pseudo-labeled data in the APG stage during the training process of the UDAA model.

In conclusion, the models for tasks at various stages act as bridge from high-resource domains to zero-resource domains, facilitating knowledge transfer. The removal of any component has an impact on the experimental results, proving their positive influence on the UDAA framework.

### 5.3.2 Task Combination

| Components | CBS | Twitter | WNUT2016 |
|---|---|---|---|
| MLMADL | **76.64** | **68.81** | **68.24** |
| NERADL | 76.59 | 68.13 | 67.62 |
| MLMADL & NERADL | 76.03 | 68.18 | 67.30 |

Table 6: Experimental results on tasks combined with DAAN.

We explore the impact of DAAN when integrated with both NER and non-NER task, represented as follows: i) MLMADL, which combines DAAN with the MLM task, ii) NERADL, which combines DAAN with the NER task, and iii) MLMADL & NERADL, which combines DAAN with the MLM task and NER task. Based on the results presented in Table 6, it is evident that the MLMADL component achieves the best results on the three datasets.

The MLMADL & NERADL strategy yielded the lowest experimental results across the three datasets. This may be attributed to the introduction of noisy data during training due to potential differences between tasks after integrating DAAN into both tasks, leading to a decrease in model performance. Therefore, in the training process of the UDAA model, we opted for the strategy of training MLM combined with DAAN. Furthermore, considering the high computational resource demand of the MRC task and its poor experimental results, we did not conduct experiments to construct MRCADL.

## 5.4 Case Study

| | |
|---|---|
| **CBS** | Astronomer [Kobie Van] with the [Johannesburg Planetarium] in [South Africa] told to [CBS News] foreign correspondent Debora [Patta] the reason for ... |
| TOF | [ Kobie Van $PER$ ] [ Johannesburg Planetarium $LOC$ ] [CBS News ORG ] [Patta PER ] |
| UDAA | [ Kobie Van $PER$ ] [ Johannesburg Planetarium $ORG$ ] [CBS News ORG ] [Patta PER ] |
| **Twitter** | Caress your phrase tenderly : it will end by smiling at you. [Anatole] [France] #amwriting #writing http://t.co/48WU4phqiO |
| TOF | [ Anatole $LOC$ ] [ France $LOC$ ] |
| UDAA | [ Anatole $PER$ ] [ France $LOC$ ] |
| **WNUT2016** | Where it all begins! An exhilarating visit to source of [Arno River] in Alto #[Casentino] http://t.co/EdZIWjVxGc |
| TOF | [ Casentino $I$ ] |
| UDAA | [ Casentino $B$ ] |

Table 7: Cases example of three datasets, where the red and blue represent correct and incorrect entities, respectively.

Table 7 presents case examples from three datasets, exhibit the performance comparison between TOF and UDAA. In the CBS dataset, the words “Kobie Van”, “CBS News”, and “Debora Patta” are correctly identified as entities in both TOF and UDAA. However, due to the influence of the word “South Africa” in the following context, TOF erroneously identifies “Johannesburg Planetarium” as the $LOC$ entity,

whereas UDAA can better understand the contextual information and correctly identify it as the $ORG$ entity.

In the Twitter and WNUT2016 datasets from the social media domain, "#" is used to highlight specific topics, which constitutes domain-specific knowledge, posing challenges for entity prediction by TOF. However, UDAA can accurately identify and classify entities. For example, in the Twitter dataset, due to the influence of "#" and"France" in the context, TOF erroneously identifies "Anatole" as the $LOC$ entity, while UDAA correctly identifies it as the $PER$ entity. In summary, the examples in Table 7 visually demonstrate the superiority of the UDAA model in distinguishing domain-specific knowledge and context understanding.

## 6 Conclusion

We propose the UDAA framework, which combines the MLM auxiliary task with the DAAN to mitigate inter-domain differences. The models for tasks at different stages serve as bridges from high-resource domains to zero-resource domains and enable knowledge transfer. Through extensive experimental evaluation using the CBS, Twitter, and WNUT2016 datasets, we demonstrate the effectiveness of UDAA in the zero-resource cross-domain NER task. In the future, we plan to improve the optimization of pseudo-data by prompt learning and explore the task where label mismatch scenarios are involved.

## 7 Acknowledgements

This work is supported by the science and technology project of State Grid Corporation of China (Grant No. 5400-202355230A-1-1-ZN).

## References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *International Conference on Computational Linguistics*.

M Saiful Bari, Shafiq Joty, and Prathyusha Jwalapuram. 2020. Zero-resource cross-lingual named entity recognition. In *Proceedings of the aaai conference on artificial intelligence*, volume 34, pages 7415–7423.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Xiaochuang Han and Jacob Eisenstein. 2019. Unsupervised domain adaptation of contextualized embeddings for sequence labeling. *arXiv preprint arXiv:1904.02817*.

Jinpeng Hu, Yaling Shen, Yang Liu, Xiang Wan, and Tsung-Hui Chang. 2022. Hero-gang neural model for named entity recognition. *arXiv preprint arXiv:2205.07177*.

Chen Jia and Yue Zhang. 2020. Multi-cell compositional lstm for ner domain adaptation. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5906–5917.

Chen Jia, Xiaobo Liang, and Yue Zhang. 2019. Cross-domain ner using cross-domain language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2464–2474.

Constantinos Karouzos, Georgios Paraskevopoulos, and Alexandros Potamianos. 2021. Udalm: Unsupervised domain adaptation through language modeling. *arXiv preprint arXiv:2104.07078*.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2019. A unified mrc framework for named entity recognition. *arXiv preprint arXiv:1910.11476*.

CCL 2024

Bill Yuchen Lin and Wei Lu. 2018. Neural adaptation layers for cross-domain named entity recognition. *arXiv preprint arXiv:1810.06368*.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2017. Adversarial multi-task learning for text classification. *arXiv preprint arXiv:1704.05742*.

Jian Liu, Yubo Chen, Kang Liu, Wei Bi, and Xiaojiang Liu. 2020a. Event extraction as machine reading comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1641–1651.

Zihan Liu, Genta Indra Winata, and Pascale Fung. 2020b. Zero-resource cross-domain named entity recognition. *arXiv preprint arXiv:2002.05923*.

Zihan Liu, Genta Indra Winata, Peng Xu, and Pascale Fung. 2020c. Coach: A coarse-to-fine approach for cross-domain slot filling. *arXiv preprint arXiv:2004.11727*.

Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, Andrea Madotto, and Pascale Fung. 2021. Crossner: Evaluating cross-domain named entity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13452–13460.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Matthew E Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. *arXiv preprint arXiv:1705.00108*.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, and Luke Zettlemoyer. 2018. Deep contextualized word representations.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. Mad-x: An adapter-based framework for multi-task cross-lingual transfer. *arXiv preprint arXiv:2005.00052*.

Jason Phang, Iacer Calixto, Phu Mon Htut, Yada Pruksachatkun, Haokun Liu, Clara Vania, Katharina Kann, and Samuel R Bowman. 2020. English intermediate-task training improves zero-shot cross-lingual transfer too. *arXiv preprint arXiv:2005.13013*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Marek Rei. 2017. Semi-supervised multitask learning for sequence labeling. *arXiv preprint arXiv:1704.07156*.

Erik F Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

Benjamin Strauss, Bethany Toma, Alan Ritter, Marie-Catherine De Marneffe, and Wei Xu. 2016. Results of the wnut16 named entity recognition shared task. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 138–144.

Tianxiang Sun, Yunfan Shao, Xiaonan Li, Pengfei Liu, Hang Yan, Xipeng Qiu, and Xuanjing Huang. 2020. Learning sparse sharing architectures for multiple tasks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8936–8943.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2016. Newsqa: A machine comprehension dataset. *arXiv preprint arXiv:1611.09830*.

Jing Wang, Mayank Kulkarni, and Daniel Preoţiuc-Pietro. 2020. Multi-domain named entity recognition with genre-aware and agnostic inference. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8476–8488.

Genta Indra Winata, Zhaojiang Lin, Jamin Shin, Zihan Liu, and Pascale Fung. 2019. Hierarchical meta-embeddings for code-switching named entity recognition. *arXiv preprint arXiv:1909.08504*.

Wei Wu, Fei Wang, Arianna Yuan, Fei Wu, and Jiwei Li. 2019. Coreference resolution as query-based span prediction. *arXiv preprint arXiv:1911.01746*.

CCL 2024

Wenhan Xiong, Jiawei Wu, Hong Wang, Vivek Kulkarni, Mo Yu, Shiyu Chang, Xiaoxiao Guo, and William Yang Wang. 2019. Tweetqa: A social media focused question answering dataset. *arXiv preprint arXiv:1907.06292*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020a. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Mengge Xue, Bowen Yu, Zhenyu Zhang, Tingwen Liu, Yue Zhang, and Bin Wang. 2020b. Coarse-to-fine pre-training for named entity recognition. *arXiv preprint arXiv:2010.08210*.

Jie Yang, Shuailong Liang, and Yue Zhang. 2018. Design challenges and misconceptions in neural sequence labeling. *arXiv preprint arXiv:1806.04470*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022a. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Jiali Zeng, Yufan Jiang, Yongjing Yin, Xu Wang, Binghuai Lin, and Yunbo Cao. 2022b. Dualner: A dual-teaching framework for zero-shot cross-lingual named entity recognition. *arXiv preprint arXiv:2211.08104*.

Guodong Zhang, Chaoqi Wang, Bowen Xu, and Roger Grosse. 2018a. Three mechanisms of weight decay regularization. *arXiv preprint arXiv:1810.12281*.

Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. 2018b. Adaptive co-attention network for named entity recognition in tweets. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Ying Zhang, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. 2021. Target-oriented fine-tuning for zero-resource named entity recognition. *arXiv preprint arXiv:2107.10523*.

Yichun Zhao, Jintao Du, Gongshen Liu, and Huijia Zhu. 2022. Transadv: A translation-based adversarial learning framework for zero-resource cross-lingual named entity recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 742–749.

Junhao Zheng, Haibin Chen, and Qianli Ma. 2022. Cross-domain named entity recognition via graph matching. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2670–2680.

CCL 2024

# Triple-view Event Hierarchy Model for Biomedical Event Representation

**Jiayi Huang**[1], **Lishuang Li**[1*], **Xueyang Qin**[1], **Yi Xiang**[1], **Jiaqi Li**[2], **Yubo Feng**[1]
[1] Dalian University of Technology, Dalian, China
[2] China Foreign Affairs University, Beijing, China
jyhuang1207@163.com, lils@dlut.edu.cn, qinxueyang@snnu.edu.cn
xiangyi_job@163.com, ljiaqi@mail.cfau.edu.cn, argamx@126.com

## Abstract

Biomedical event representation can be applied to various language tasks. A biomedical event often involves multiple biomedical entities and trigger words, and the event structure is complex. However, existing research on event representation mainly focuses on the general domain. If models from the general domain are directly transferred to biomedical event representation, the results may not be satisfactory. We argue that biomedical events can be divided into three hierarchies, each containing unique feature information. Therefore, we propose the Triple-views Event Hierarchy Model (TEHM) to enhance the quality of biomedical event representation. TEHM extracts feature information from three different views and integrates them. Specifically, due to the complexity of biomedical events, We propose the Trigger-aware Aggregator module to handle complex units within biomedical events. Additionally, we annotate two similarity task datasets in the biomedical domain using annotation standards from the general domain. Extensive experiments demonstrate that TEHM achieves state-of-the-art performance on biomedical similarity tasks and biomedical event casual relation extraction.

## 1 Introduction

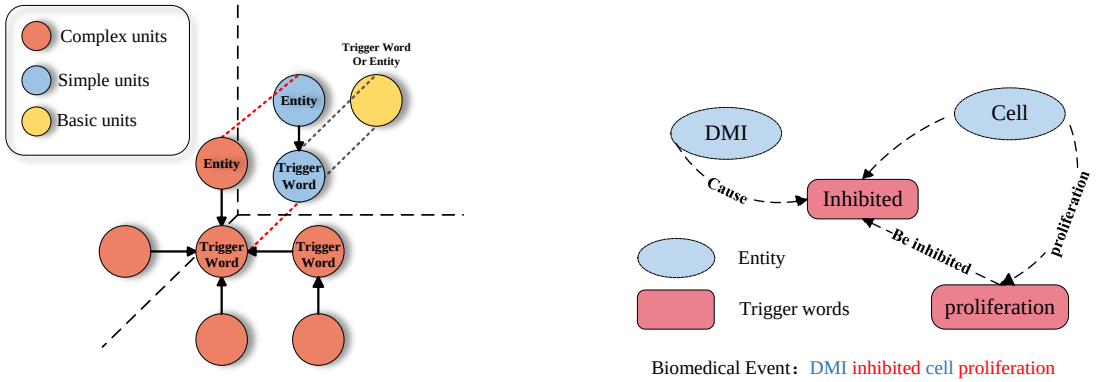
Biomedical events consist of trigger words, biomedical entities, and specific actions of entities (Frisoni et al., 2022). A complex biomedical event often contains multiple simple biomedical events, making the structure of event texts quite complex. Biomedical event representation refers to the process of converting biomedical event texts into machine-readable formats. Event representation is applied in tasks such as event prediction, event relation extraction, and event information extraction (Hwang et al., 2022), which can improve the performance of related tasks. Therefore, obtaining a representation of biomedical events is necessary.

Currently, in the general domain, the works on event representation focus on simple events, such as "He was frightened football" which typically follow a subject-verb-object structure. These works usually rely on pre-trained language models (PLMs) for text encoding (Gao et al., 2022). However, biomedical events involve multiple biomedical entities, and the event structure is not a simple subject-verb-object format. For example, in the biomedical event "cooperation among NF-kappa B-, AP-1- and NF-AT-binding sequences is required for induction of the GM-CSF gene through PKC- and Ca2+- signaling pathways downstream of T-cell activation" there are multiple verbs and the relationships between verbs are nested within each other (Weber et al., 2018). If we directly transfer models from the general domain for biomedical event representation, we cannot extract and integrate the structural features of biomedical events due to their complexity. Currently, existing works lack biomedical event representation models capable of handling complex structures.

We argue that although biomedical events are complex, they can be categorized into three hierarchies. As illustrated in Figure 1(a), we divide a complex biomedical event into complex units, simple units, and basic units. Hierarchically organizing biomedical events and deriving biomedical event representations from these three hierarchies is an effective approach. Therefore, based on this perspective, we

*Corresponding author: Lishuang Li(lils@dlut.edu.cn)
©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1136-1147, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China




(a) Triple-views Event Hierarchy

(b) Complex units: Trigger-aware Graph

Figure 1: Figure 1(a) depicts the structure of a biomedical event. Trigger words and entities belong to the basic unit. A simple unit is an event that contains only one trigger word. A complex unit involves multiple trigger words. Figure 1(b) illustrates the event graph structure we constructed for processing complex units with graph neural networks, centered around trigger words.

propose the Triple-views Event Hierarchy Model (TEHM). The TEHM learns triple hierarchical features of biomedical events from three views. From the complex units view, we propose the Trigger-aware Aggregator module, which focuses on extracting features from complex units with the trigger word as the core. It is based on the architecture of graph neural networks and has been improved to better suit the characteristics of biomedical events for feature extraction. From the simple units view, we introduce the Context-aware Intensifier module, which integrates context information through prompt learning to explore implicit interaction features within simple events. From the basic units view, we present the Entity Relation Integrator module, which obtains feature encoding matrices for biomedical entities. This module constructs an entity relation network and employs the algorithm proposed in this paper BioER to train entity feature matrices.

After obtaining event representations, how to evaluate the quality of event representations is also a problem that needs to be addressed currently. Due to the significant differences in datasets between the general domain (which consists of simple events) and the biomedical domain, there is currently a notable lack of effective evaluation tasks to directly validate the quality of biomedical event representations. To address this issue, we adopt the same annotation standards (Weber et al., 2018; Kartsaklis and Sadrzadeh, 2014) as those used for the tasks in the general domain and manually annotate two validation tasks relevant to assessing the quality of biomedical event representations: Bio Hard Similarity Task and Bio Transtive Sentence Similarity Task. Additionally, we validate the effectiveness of biomedical event representation on standard biomedical event causal relation extraction datasets.

After extensive experiments, TEHM achieves improvements of 25.9% (accuracy) in Bio Hard Similarity and 0.045 (Spearman correlation coefficient) in Bio Transitive Sentence Similarity. Additionally, it improves 3.0% in the F1 score on Hahn Powell's Dataset for biomedical event causal relation extraction, reaching the current SOTA.

In summary, the contributions of this paper can be summarized as follows:

- To the best of our knowledge, we are the first to propose the task of biomedical event representation. We partition complex biomedical events into three hierarchies and propose the Triple-views Event Hierarchy Model (TEHM) for biomedical event representation. TEHM extracts features from three views, integrates them, and produces the final representation of biomedical events.

- Based on the characteristics of complex units in biomedical events, we propose a Trigger-aware Aggregator module with the trigger word as the core. This module is designed to explore the relational features within complex biomedical events.

- We construct a dataset for evaluating the quality of biomedical event representation. We will make it publicly available, laying the foundation for further research. Simultaneously, we achieve SOTA in similarity tasks and biomedical event relation causality extraction.

## 2 Triple-view Event Hierarchy Model(TEHM)

Figure 2 illustrates the framework of the TEHM model. The input to the TEHM model is the event text, which is simultaneously processed by the feature extraction modules of TEHM to obtain features from three different views of the event. These feature vectors are then concatenated to obtain the representation of the final biomedical event. TEHM employs a contrastive learning approach to training the model, which is a form of unsupervised learning.

### 2.1 Problem definition

**Event Representation.** Our goal is to enable the model to learn features of biomedical events, obtain embeddings for these events, and subsequently investigate the model's event representation capability. Specifically, given a similar event pairs $\left\{v_i^+, v_j^+\right\}$ and a dissimilar event pairs $\left\{v_i^-, v_j^-\right\}$, along with a similarity evaluation function $S$, where $v$ refers to event representation. When evaluated using the function $S$, the great event representations should satisfy equation 1:

$$S\left\{v_i^+, v_j^+\right\} > S\left\{v_i^-, v_j^-\right\} \tag{1}$$

**Biomedical Event Hierarchy.** As shown in Figure 1(a), this paper defines three hierarchical units for biomedical events: basic units (which refer to **trigger words or entity elements**), simple units ( which refer to **one simple event** consisting of a trigger word and entity elements), and complex units (which refer to a complex event consisting of **two or more simple events**)

### 2.2 Trigger-aware Aggregator (TAA)

For the complex units, we model it as a Trigger-aware graph (Figure 1 (b)), utilizing the structured information from the bioT2E dataset (Frisoni et al., 2022). We represent the Trigger-aware graph as: $\mathcal{G}_i = \{\nu_i, \varepsilon_i\}$, where $\nu_i$, $\varepsilon_i$ represents a set of nodes and edges for $\mathcal{G}_i$ respectively. A set of $n$ distinct trigger words and entities is represented as $\nu_i = \{w_{i,1}, w_{i,2}, ..., w_{i,n-1}, w_{i,n}\}$ . $\varepsilon_i = \{l_{i,1}, l_{i,2}, ..., l_{i,m-1}, l_{i,m}\}$ represents a set of $m$ directed edges. Based on the structural information, we can directly establish edges between nodes in the Trigger-aware graph. The motivation behind our proposed Trigger-aware graph structure is as follows: for complex units, as the complexity of event texts increases and the number of entities and trigger words rises, directly modeling the event texts is not conducive to extracting implicit entity relationship features. We consider that different biomedical entities relate to each other due to trigger words. Therefore, using the trigger words as central hubs can uncover hidden relationships among different events that are not directly related.

We initialize the representation for nodes using BERT. Given a node $\nu_{i,n}$ from the event graph $\mathcal{G}_i$, we employ the Dropout technique by feeding the same input $\nu_{i,j}$ into BERT twice, obtaining $\vec{h}_{i,j}$ and $\vec{h}_{i,j}^+$, where $\vec{h}_{i,j}, \vec{h}_{i,j}^+ \in R^F$, and $F$ represents the feature dimension of the node input. It is important to note that different Dropout masks are applied for each of the two inputs:

$$\vec{h}_{i,j} = f_\theta\left(w_{i,j}, \phi_1\right), \vec{h}_{i,j}^+ = f_\theta\left(w_{i,j}, \phi_2\right), \tag{2}$$

where $\phi_1$ and $\phi_2$ represent two independent random masks, and $f_\theta$ represents the BERT encoding. We can obtain different vector representations for the same graph node by using different masks.

The subsequent steps involve facilitating information exchanges between nodes by utilizing the central node. For the Trigger-aware graph, we input a set of node features $h_i = \left\{\vec{h}_{i,1}, \vec{h}_{i,2}, ..., \vec{h}_{i,n}\right\}$, where $n$ represents the number of nodes in the event graph. To enhance the importance of the central node, we incorporate the dimension information of the node itself in the aggregation layer of the graph. As illustrated

CCL 2024

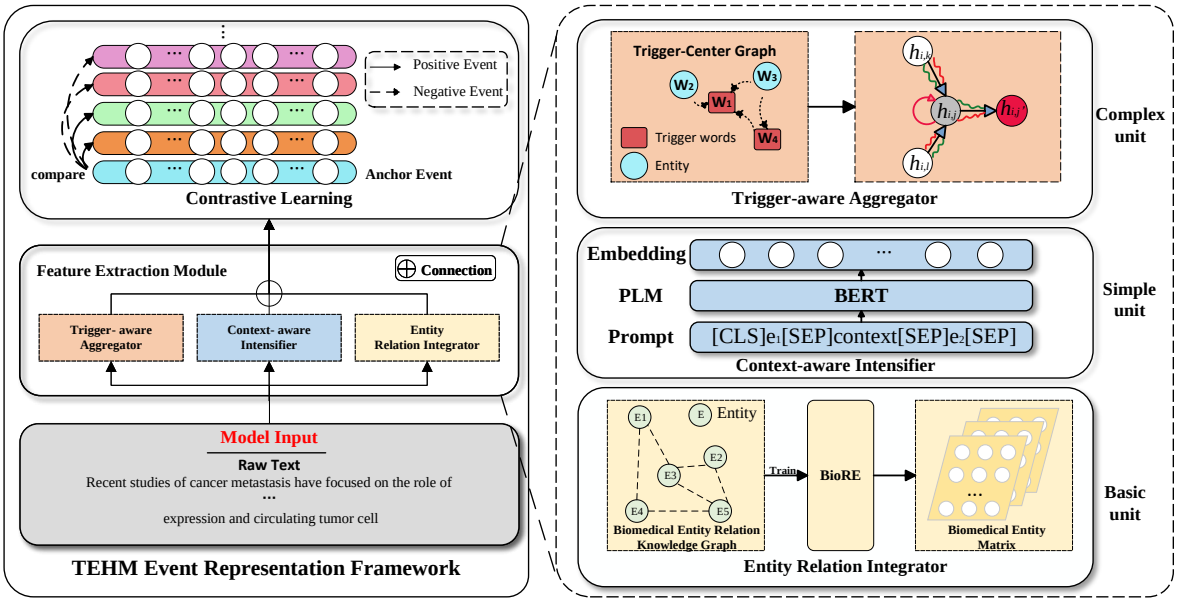



Figure 2: The overall framework of the TEHM model (left) and details the processing of biomedical events from triple views (right). After unsupervised learning, we obtain a well-trained TEHM model. The model's output serves as the representation of biomedical events.

in Figure 2, when $\vec{h}_{i,j}$ passes through the graph attention layer, it produces $h_i^{'} = \left\{\vec{h}_{i,1}^{'}, \vec{h}_{i,2}^{'}, ...\vec{h}_{i,n}^{'}\right\}$ where $\vec{h}_{i,j}^{'} \in R^{F'}$, and $F^{'}$ represents the feature dimension of the node output:

$$\vec{h}_{i,j}^{'} = sigmoid\left(\sum_{k \in N_i} \alpha_{ik} W_h \vec{h}_{i,k} + W_h \vec{h}_{i,j}\right), \tag{3}$$

where, $N_{ij}$ represents the neighborhood formed by all nodes adjacent to $w_{i,j}$, and $W_h$ is the feature matrix with dimensions $F \times F^{'}$. $\alpha_{ik}$ represents the attention coefficient for information exchange between node $h_{i,j}$ and node $h_{i,k}$, indicating the importance of node $h_{i,k}$ corresponding to node $h_{i,j}$. The calculation formula in this paper is as follows:

$$\alpha_{ik} = softmax\left(atten_{ik}\right) = \frac{exp\left(atten_{ik}\right)}{\sum_{j \in N_{ij}} exp\left(atten_{ik}\right)}, \tag{4}$$

where $atten_{ij}$ represents attention coefficient. Its calculation formula is shown as Equation 5:

$$atten_{ik} = \frac{exp\left\{\sigma\left(\vec{o}^T\left[W_a \vec{h}_{i,j} \oplus W_a \vec{h}_{i,k}\right]\right)\right\}}{\sum_{k \in N_i} exp\left\{\sigma\left(\vec{o}^T\left[W_a \vec{h}_{i,j} \oplus W_a \vec{h}_{i,k}\right]\right)\right\}}, \tag{5}$$

where $\oplus$ represents the concatenation of two matrices on both sides, $\sigma$ represents LeakyRelu avtivate function, $\vec{o}$ is a weight vector with dimensions $\vec{o} \in R^{2F^{'}}$, $W_a$ is the training weight matrix. The multiplication operation between them results in a scalar value. The calculation yields $atten_{ik}$. Substituting it into equations 4 and 3 obtains the enhanced node representation after TAA.

Finally, we convert the node representations into integrated vectors $\vec{y}_i$ (likewise, using $\vec{h}_{i,j}^{+}$ as input into Equations (3 - 6) yields $\vec{y}_i^{+}$) the equation as follow:

$$\vec{y}_i = \frac{1}{n} \sum_{j \in h_i^{'}} \left[\vec{h}_{i,j}^{'}\right]^T, \qquad \vec{y}_i^{+} = \frac{1}{n} \sum_{j \in h_i^{'+}} \left[\vec{h}_{i,j}^{'+}\right]^T, \tag{6}$$

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1136-1147, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

where, $n$ denotes the nodes number of $\mathcal{G}_i$.

### 2.3 Context-aware Intensifier (CAI)

This part of the model is designed to handle simple units of biomedical events. We find that when people need to understand an event, they usually need to combine the context to grasp the meaning conveyed by the event. The context becomes even more important when we need to do reasoning. Therefore, injecting contextual information allows the model to uncover implicit event relationships and extract features of event relationships. This is also why TEHM can be applied to event relationship extraction. Additionally, annotation in biomedical event relationship data is expensive and challenging. However, our model is trained in unsupervised paradigms, addressing the annotation issue.

We place the contextual information between the head event $e_1$ and the tail event $e_2$. This effectively informs the model about two events along with the contextual information between them, enabling the model to explore their implicit relationships. We construct the Prompt as shown in the formula:

$$Prompt : [CLS]\, e_1\, [SEP]\, context\, [SEP]\, e_2\, [SEP]\,, \tag{7}$$

This Prompt helps the model learn event relationship information, enhances the features of the events themselves, and strengthens the capability of TEHM for event relationship extraction.

To obtain representation and its positive example, we also employ the Dropout technique to input the above template $p_i$ into BERT twice:

$$\vec{z}_i = f_\theta\left(p_i, \phi_1\right), \vec{z}_i^{+} = f_\theta\left(p_i, \phi_2\right). \tag{8}$$

### 2.4 Entity Relation Integrator (ELI)

**Algorithm 1** BioER

**Input:** Training set $S = \{h, r, t\}$, entities and relations sets. $E$ and $R$, embeddings dim. k

**initialize** $r \leftarrow$ uniform$(-\frac{6}{\sqrt{k}}, \frac{6}{\sqrt{k}})$ for each $r \in R$
  $r \leftarrow r/\|\mathrm{r}\|$ for each $r \in R$
  $e \leftarrow$ uniform $(-\frac{6}{\sqrt{k}}, \frac{6}{\sqrt{k}})$ for each entity $e \in E$
**loop**
  $e \leftarrow e/\|\mathrm{e}\|$ for each $e \in E$
  $S_{batch} \leftarrow$ sample($S$,b)//sample a minibatch of size b
  $T_{batch} \leftarrow \emptyset$//initialize the set of pairs of triplets
  **for** $(h, r, t) \in S_{batch}$ **do**
    $(h^{'}, r^{'}, t^{'}) \leftarrow$sample$(S^{'}_{(h,r,t)})$ // sample a corrupted triplet
    $T_{batch} \leftarrow T_{batch} \bigcup \left\{((h, r, t), (h^{'}, r^{'}, t^{'}))\right\}$
  **end for**
  Updata embedding $\sum_{((h,r,t),(h^{'},r^{'},t^{'}))\in T_{batch}} -log\frac{d(h+r,t)}{d(h+r,t)+d(h^{'}+r^{'}-t^{'})}$
**end loop**

To extract the features of base units, we utilize entity relationship information provided by the Unified Medical Language System (UMLS) to establish a biomedical entity-relationship network. This paper posits that implicit entity relationship features can enhance event representation, thus facilitating the acquisition of high-quality event representations. As shown in Algorithm 1, to obtain knowledge embeddings for entity relations, we propose a new algorithm called BioER. Our training objective is to separate positive and negative examples as much as possible. The distance calculation formula is equation 9:

$$d(h + r, t) = \|h + r - t\|, \tag{9}$$

where the distance formula $d(h + r, t)$ can take either the L1 or L2 norm.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1136-1147, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

The core of the algorithm is to make the distance $d(h+r,t)$ of positive examples approach zero, while the distance $d(h^{'}+r^{'}-t^{'})$ of negative examples approach infinity. However, The traditional TransE's loss function is based on Hinge loss, which requires an additional margin distance parameter. This loss function cannot drive the negative value of $d(h^{'}+r^{'}-t^{'})$ to approach infinity, and its convergence heavily depends on the margin distance. Therefore, we improve the loss function here to address these issues.

We train a well-performing BioER model Using the aforementioned algorithm. Subsequently, we utilize BioER to encode the biomedical entities within the events, obtaining a biomedical entity matrix, as illustrated in Figure 2. When inputting corresponding biomedical entities, we search for the corresponding vectors in the biomedical entity matrix. Then, we perform average pooling on the entity vectors to obtain the feature vector $\vec{r}_i$ for the biomedical event $i$.

### 2.5 Model Training Objection

Finally, we obtain the final representation $\vec{v}_i$ of the biomedical event along with its positive sample $\vec{v}_i^{+}$ by performing the following operations on $\vec{z}_i$, $\vec{y}_i$, $\vec{r}_i$, and $\vec{z}_i^{+}$, $\vec{y}_i^{+}$, $\vec{r}_i^{+}$. As shown in Equation 10:

$$\vec{v}_i = \vec{y}_i \oplus \vec{z}_i \oplus \vec{r}_i, \qquad \vec{v}_i^{+} = \vec{y}_i^{+} \oplus \vec{z}_i^{+} \oplus \vec{r}_i^{+}. \tag{10}$$

Contrastive learning allows the model to better capture subtle features of positive and negative examples. Therefore, similar to (Gao et al., 2022), we adopt the multi-positive-sample InfoNCE(Gao et al., 2022) loss function for training the model. When presenting a set containing events $x_i$, the loss function is defined as follows:

$$\mathcal{L} = \sum_{a \in A_i} -\log \frac{g\left(\vec{v}_i, \vec{v}_{ia}^{+}\right)}{g\left(\vec{v}_i, \vec{v}_{ia}^{+}\right) + \sum_{k \in M_i} g\left(\vec{v}_i, \vec{v}_k\right)}, \tag{11}$$

where A(i) and M(i) respectively denote the sets of positive and negative examples for the biomedical event $x_i$. We use $A_i = \left\{\vec{v}_i^{+1}, \vec{v}_i^{+2}\right\}$. Here, $k$ represents the index of negatives within a sample batch, and $g$ is defined as: $g(\vec{v}_i, \vec{v}_k) = exp(\vec{v}_i^{T}\vec{v}_k/\tau)$, where $\tau$ is a hyperparameter known as temperature.

## 3 Biomedical Event Representation Evaluation Dataset

**BioT2E** (Frisoni et al., 2022): This dataset provides structured biomedical event text information, including event triggers, biomedical entities, and entity-specific attributes. After filtering out texts without events and single-trigger events, we obtain a total of 36,662 biomedical events.

- **Bio Hard Similarity**: Similar to the hard similarity task(Weber et al., 2018), we have one annotator create similar/dissimilar pairs from BioT2E, while three different annotators give the similarity/dissimilarity rankings. We keep pairs where the annotators agree completely. Finally, we select 240 event pairs from BioT2E. It includes pairs of biomedical events that are semantically similar but have minimal lexical overlaps, as well as pairs that are semantically different but have significant lexical overlaps. For example, "accelerated resistance TGF-beta cells | speedy resistibility TGF-beta cells | accelerated resistance TGF-beta cells | accelerated activation TGF-beta cells". "|" indicates the separator. We use accuracy as the metric to evaluate this task, measuring the accuracy of the model in correctly identifying similar event pairs with higher cosine similarity representations compared to dissimilar event pairs.

- **Bio Transtive Sentence Similarity**: We select 120 event pairs from BioT2E and use standard annotations from the general domain (Kartsaklis and Sadrzadeh, 2014) for this task. Each pair of data is manually annotated with a similarity score ranging from 1 to 7. For example, "leading secretion IL-8 induce transcription IL-8 | leading secretion IL-8 cause tumor apoptosis | 4.1". The final score, 4.1 in this case, represents the similarity score. This dataset comprises a total of 120 similar data instances. A higher score indicates greater similarity between the events. We compute the Spearman correlation coefficient ($\rho \in [-1, 1]$) between the similarity scores predicted by the model and the similarity scores manually annotated.

CCL 2024

**Hahn-Powell's**(Hahn-Powell et al., 2016): The dataset is about extracting causal relations in biomedical events, and it's currently one of the most commonly used datasets for biomedical event causal relation extraction. Given two events in the text and determining whether there is a causal relationship between the two events, we compute the precision (P), recall (R), and F1-score (F) for this task.

## 4 Experiments

**Implementation Details.** During the training process, the batch size is set to 32, and our model uses the Adam optimizer. The learning rate for the Adam optimizer of the Context-aware Intensifier is set to 3e-5, while the Adam optimizer used for the Trigger-aware Aggregator has a learning rate of 3e-7. The temperature parameter $\tau$ is set to 0.3.

### 4.1 Main results

We compare the TEHM model with several PLMs in the field of event representation: BERT (Kenton and Toutanova, 2019), Roberta (Liu et al., 2019), BioBERT (Lee et al., 2020), and T5 (Raffel et al., 2020). We also compare two models that use contrastive learning for event representation: Simcse (Chen and He, 2021), SWCC (Gao et al., 2022) and PromptCL (Feng et al., 2023). Performance of various models in Bio Hard Similarity and Bio Transitive Sentence Similarity is reported in Table 1.

The experimental results indicate that our TEHM demonstrates superior performance in similarity tasks for biomedical event representation. It not only outperforms current large-scale language models but also surpasses the SOTA model PromptCL (which be used in the general domain) by 26.1% in the Bio Hard Similarity task.

| **Model** | **Bio Hard Similarity** Acc.% | **Bio Transitive Sentence Similarity**($\rho$) |
|---|---|---|
| BERT(Kenton and Toutanova, 2019) | 42.5 | -0.011 |
| BioBERT(Lee et al., 2020) | 41.7 | 0.148 |
| RoBERTa(Liu et al., 2019) | 38.3 | 0.525 |
| T5(Raffel et al., 2020) | 50.8 | 0.284 |
| SimCSE(Chen and He, 2021) | 41.7 | 0.643 |
| SWCC(Gao et al., 2022) | 49.2 | 0.412 |
| PromptCL(Feng et al., 2023) | 50.6 | 0.410 |
| **TEHM(Our)** | **76.7** | **0.688** |

Table 1: Performance in Bio Hard Similarity task and Bio Transitive Sentence Similarity task

These results adequately prove the effectiveness of the TEHM model designed for the structure of biomedical events in improving the quality of biomedical event representation. We argue that there are three reasons why TEHM achieves state-of-the-art performance in biomedical event representation:

- **From a holistic perspective:** We partition biomedical events into three hierarchies and enable TEHM to capture features from three views, then merge them. This helps TEHM to comprehend events at different hierarchies, thereby enhancing the quality of event representation.
- **From a local perspective:** Due to the complexity of biomedical events, we propose the Trigger-aware Aggregator module. It focuses on the trigger word to extract the relationships between various entities and the trigger word within complex events. This architecture enables the model to emphasize the importance of trigger words when learning event representations.
- **From an external knowledge perspective:** We integrate context to capture implicit relationship features within events. Here's an intuitive explanation: rich contextual information aids our understanding of event text when we read it, enabling us to analyze the inherent relationships among multiple simple events within complex events. Additionally, the relationships between biomedical

CCL 2024

entities that we integrate are specific to biomedical entities, helping the model understand events from the most basic hierarchies.

### 4.2 Biomedical Event Casual Relation Extraction

To further validate the representation capability of TEHM in biomedical events, we apply TEHM to the SOTA biomedical causal relation extraction model, MKFN. Table 2 demonstrates the superior performance of TEHM in biomedical event casual relation extraction on Hahn Powell's dataset. This is because the given data text typically contains multiple events. Handling such complex event texts is precisely the problem that the TEHM model aims to address. TEHM can perceive nearby entities and other trigger words centered around the trigger word while incorporating contextual information to explore implicit causal relationships within the text. This enhancement results in a 3.0% performance improvement when integrating event representations into the MKFN on Hahn Powell's dataset.

| Model | P(%) | R(%) | F1(%) |
|---|---|---|---|
| LSTM (Akkasi and Moens, 2021) | 40.0 | 25.0 | 31.0 |
| SVM+L1 (Akkasi and Moens, 2021) | 54.0 | 35.0 | 43.0 |
| Graph state LSTM (Akkasi and Moens, 2021) | 57.0 | 45.0 | 50.0 |
| Multiview-CNN (Akkasi and Moens, 2021) | 60.0 | 44.0 | 52.0 |
| BiLSTM+ATTENTION (Akkasi and Moens, 2021) | 54.0 | 49.0 | 52.0 |
| BERT (Akkasi and Moens, 2021) | 53.7 | 51.2 | 52.0 |
| BioBERT+CNN (Liang et al., 2022) | 63.0 | 53.0 | 58.0 |
| MKFN (Hao et al., 2023) | 69.7 | 58.9 | 63.9 |
| **MKFN+TEHM(Our)** | **72.4** | **62.2** | **66.9** |

Table 2: Biomedical Event Casual Relation Extraction on Hahn Powell's Dataset

### 4.3 Ablation Experiments

To study the influence of different modules in TEHM on its overall performance, we conduct a series of ablation experiments on Bio Hard similarity and Bio Transtive Sentence similarity. We can draw the following conclusions from Table 3:

| Model | Bio Hard similarity Acc.% | Bio Transitive sentence similarity($\rho$) |
|---|---|---|
| **TEHM** | 76.7 | 0.688 |
| –w/o Trigger-aware Aggregator | 54.2 | 0.664 |
| –w/o Event-Context Intensifier | 65.8 | 0.648 |
| –w/o Entity Relation Integrator | 75.8 | 0.688 |
| –w/o Prompt | 55.0 | 0.661 |

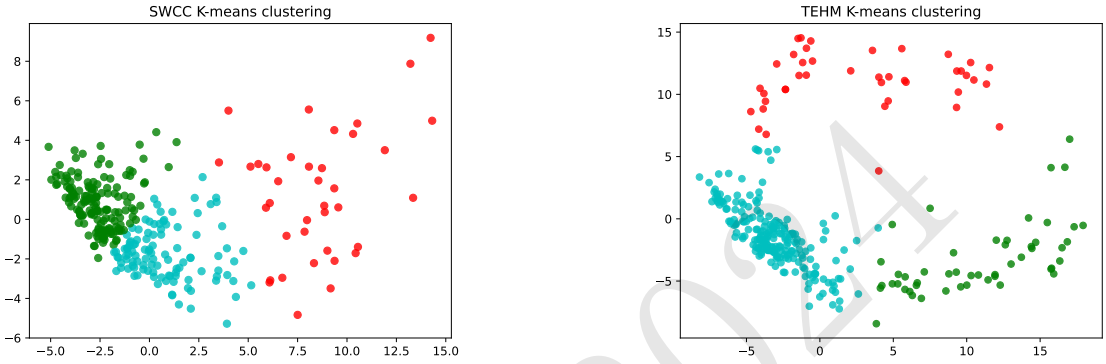
Table 3: Ablation Study For TEHM On Similarity Tasks

- We remove the Trigger-aware Aggregator (TAA) module, resulting in a performance decrease of 22.5% in Bio Hard similarity and 0.024 in Bio Transtive Sentence similarity. Our findings suggest that the TAA plays a significant role in the Bio Hard similarity task, indicating its effectiveness in capturing subtle semantic features. This capability allows the model to distinguish events with high word overlaps but different semantics, mapping them to different vector spaces.
- We remove the Context-aware Intensifier (CAI) module, resulting in a significant decrease in performance by 0.04 for Bio Transitive Sentence similarity, which is the largest decrease among all ablation experiments. We argue that continuous event text contents are beneficial for enhancing performance in this task. This is because the task requires the model to understand the overall semantic meaning of events and calculate the Spearman correlation coefficient.

CCL 2024

- We remove the Entity Relation Integrator (ELI) module, resulting in a 0.9% decrease in performance for Bio Hard similarity and no decline in performance for Bio Transtive Sentence similarity. This is consistent with our aforementioned analysis, indicating that Bio Transtive Sentence similarity task requires the model to comprehend events from the whole event. The entity-relationship matrix only provides entity relationship information, aiding in distinguishing subtle semantic features of events.
- To investigate the impact of the Prompt we proposed, we conduct experiments without connecting context information. The results indicate a decrease in performance for Bio Hard similarity and Bio Transtive Sentence similarity tasks by 21.7% and 0.027, respectively. This suggests that injecting context information helps the model better understand the semantic information of biomedical events both at a detailed and holistic level.

### 4.4 Visualization of Learned Representations




(a) SWCC (b) TEHM

Figure 3: K-means clustering results of biomedical events

We randomly select 300 biomedical events and encode them using both the SWCC and TEHM. Subsequently, we apply Principal Component Analysis (PCA) to map the obtained vectors to a two-dimensional space. Using an unsupervised approach, we perform cluster analysis on the biomedical events with a specified cluster count of 3. The results in Figure 3 show that the vectors encoded by the SWCC lack effective clustering. We also report the clustering performance metric for both the TEHM and the SWCC model (Table 4).

| Model | Sil | CH | DBI |
|---|---|---|---|
| SWCC | 0.424 | 341.33 | 0.811 |
| TEHM | 0.623 | 408.4 | 0.610 |

Table 4: Clustering performance metrix

The evaluation involves computing three key metrics: Silhouette Score (Sil), Calinski-Harabasz Score (CH), and Davies Bouldin Score (DB). Biomedical event representations encoded by the TEHM exhibit superior performance in clustering experiments. The cluster analysis demonstrates that the biomedical event representations by the TEHM effectively bring semantically similar events closer in vector space and further separate semantically distinct biomedical events.

### 4.5 Case Study

To further analyze the performance of TEHM in learning biomedical event representations, we conduct a case study. Table 5 provides the cosine similarity of the vectors of biomedical event pairs encoded by BERT, SWCC, and TEHM.

To illustrate, we take the third and fourth as examples. In the pair of biomedical events, “transformation mammary epithelial cells” and “conversion breast epithelial cells”, which are semantically similar but have fewer overlapping words, TEHM encodes them with a cosine similarity higher than BERT by 0.0739. This case study highlights TEHM’s ability to capture more accurate and contextually meaningful

CCL 2024

event representations. In another pair "induced production TNF-alpha" and "induced production cancer", although these events have high word overlap, they are fundamentally different because TNF-alpha is a biologically active cytokine, while "cancer" refers to a disease. However, while BERT encodes these events with a cosine similarity score as high as 0.9490, TEHM yields a score of only 0.2106. Clearly, 0.2106 is more semantically meaningful.

| Event Pair | Ground Truth | BERT | SWCC | TEHM |
|---|---|---|---|---|
| induces B box localization<br>leads B box positioning | Similar | 0.6905 | 0.6562 | 0.8245 |
| induces B box localization<br>induces B box expression | Dissimilar | 0.9570 | 0.6217 | 0.1915 |
| transformation of mammary epithelial cells<br>conversion breast epithelial cells | Similar | 0.9048 | 0.8582 | 0.9833 |
| induced production of TNF-alpha<br>induced production cancer | Dissimilar | 0.9490 | 0.4780 | 0.2106 |

Table 5: Cosine Similarity Scores Case Study: We analyze **cosine similarity scores** for biomedical events encoded by different models.

## 5 Related Work

Due to the significant role of event representations in downstream applications, various methods for event representation have been introduced in current research (Zheng et al., 2024). Long Short-Term Memory (LSTM) (Huang et al., 2015) effectively integrates the sequential information of events into a unified model. Additionally, the fusion of external knowledge in jointly trained models provides a promising direction for capturing potential relationships among events (Ding et al., 2019). It's worth noting that the application of contrastive learning techniques in the field of computer vision, such as Deep InfoMax (Hjelm et al., 2019), MoCo (Long et al., 2024), and SimCSE (Chen and He, 2021), has significantly improved the performance of unsupervised image classification, even surpassing supervised methods. SWCC(Gao et al., 2022) and PromptCL (Feng et al., 2023) both use contrastive learning architecture. PromptCL has achieved SOTA performance in event representation. Recent studies (Gao et al., 2022; Wu et al., 2021) have introduced dropout noise as a data augmentation technique, highlighting its effectiveness in NLP tasks compared to many traditional augmentation methods.

## 6 Conclusion

We introduce the task of biomedical event representation for the first time and establish the evaluation tasks for biomedical event representation. We will make the relevant datasets publicly available for researchers to conduct further studies. Due to the complexity of biomedical events, we propose a hierarchical structure for biomedical events and design the TEHM model for biomedical event representation. TEHM is capable of obtaining features of biomedical events from three perspectives and capturing subtle semantic features through contrastive learning. Particularly, based on the characteristics of biomedical events, we propose a Trigger-aware Aggregator to extract relational features within complex biomedical events. Additionally, we inject context information into the model using Prompt learning. Experimental results demonstrate that we achieve state-of-the-art performance in biomedical similarity tasks. We also validate the effectiveness of event representation in improving task performance in biomedical causal relation extraction, which also achieves state-of-the-art performance.

## Acknowledgements

This work is supported by grant from the National Natural Science Foundation of China (No. 62076048).

## References

Abbas Akkasi and Mari-Francine Moens. 2021. Causal relationship extraction from biomedical text using deep neural models: A comprehensive survey. *Journal of biomedical informatics*, 119:103820.

Xinlei Chen and Kaiming He. 2021. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758.

Xiao Ding, Kuo Liao, Ting Liu, Zhongyang Li, and Junwen Duan. 2019. Event representation learning enhanced with external commonsense knowledge. pages 4893–4902.

Yubo Feng, Lishuang Li, Yi Xiang, and Xueyang Qin. 2023. PromptCL: Improving event representation via prompt template and contrastive learning. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 261–272. Springer.

Giacomo Frisoni, Gianluca Moro, and Lorenzo Balzani. 2022. Text-to-text extraction and verbalization of biomedical event graphs. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 2692–2710. International Committee on Computational Linguistics.

Jun Gao, Wei Wang, Changlong Yu, Huan Zhao, Wilfred Ng, and Ruifeng Xu. 2022. Improving event representation via simultaneous weakly supervised contrastive learning and clustering. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3036–3049. Association for Computational Linguistics.

Gus Hahn-Powell, Dane Bell, Marco Antonio Valenzuela-Escárcega, and Mihai Surdeanu. 2016. This before that: Causal precedence in the biomedical domain. In Kevin Bretonnel Cohen, Dina Demner-Fushman, Sophia Ananiadou, and Jun'ichi Tsujii, editors, *Proceedings of the 15th Workshop on Biomedical Natural Language Processing, BioNLP@ACL 2016, Berlin, Germany, August 12, 2016*, pages 146–155. Association for Computational Linguistics.

Jing Hao, Lishuang Li, and Xueyang Qin. 2023. Biomedical causal relation extraction via data augmentation and multi-source knowledge fusion. In Xingpeng Jiang, Haiying Wang, Reda Alhajj, Xiaohua Hu, Felix Engel, Mufti Mahmud, Nadia Pisanti, Xuefeng Cui, and Hong Song, editors, *IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2023, Istanbul, Turkiye, December 5-8, 2023*, pages 1112–1117. IEEE.

R. Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Philip Bachman, Adam Trischler, and Yoshua Bengio. 2019. Learning deep representations by mutual information estimation and maximization.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991.

EunJeong Hwang, Jay-Yoon Lee, Tianyi Yang, Dhruvesh Patel, Dongxu Zhang, and Andrew McCallum. 2022. Event-event relation extraction using probabilistic box embedding. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 235–244. Association for Computational Linguistics.

Dimitri Kartsaklis and Mehrnoosh Sadrzadeh. 2014. A study of entanglement in a categorical framework of natural language. 172:249–261.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Zhengzhong Liang, Enrique Noriega-Atala, Clayton Morrison, and Mihai Surdeanu. 2022. Low resource causal event detection from biomedical literature. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 252–263.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Xianzhong Long, Han Du, and Yun Li. 2024. Two momentum contrast in triplet for unsupervised visual representation learning. volume 83, pages 10467–10480.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Noah Weber, Niranjan Balasubramanian, and Nathanael Chambers. 2018. Event representations with tensor-based compositions. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 4946–4953. AAAI Press.

Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, Tie-Yan Liu, et al. 2021. R-drop: Regularized dropout for neural networks. *Advances in Neural Information Processing Systems*, 34:10890–10905.

Jianming Zheng, Fei Cai, Jun Liu, Yanxiang Ling, and Honghui Chen. 2024. Multistructure contrastive learning for pretraining event representation. *IEEE Trans. Neural Networks Learn. Syst.*, 35(1):842–854.

CCL 2024

# DialectMoE: An End-to-End Multi-Dialect Speech Recognition Model with Mixture-of-Experts

**Jie Zhou[1,2], Shengxiang Gao[*1,2], Zhengtao Yu[1,2], Ling Dong[1,2], Wenjun Wang[1,2]**
1. Faculty of Information Engineering and Automation,
Kunming University of Science and Technology, Kunming, Yunnan, 650500, China
2. Yunnan Key Laboratory of Artificial Intelligence,
Kunming University of Science and Technology, Kunming, Yunnan, 650500, China
zhoujay@stu.kust.edu.cn, gaoshengxiang.yn@foxmail.com, ztyu@hotmail.com,
ling.dong@kust.edu.cn, 20203104003@stu.kust.edu.cn

## Abstract

Dialect speech recognition has always been one of the challenges in Automatic Speech Recognition (ASR) systems. While lots of ASR systems perform well in Mandarin, their performance significantly drops when handling dialect speech. This is mainly due to the obvious differences between dialects and Mandarin in pronunciation and the data scarcity of dialect speech. In this paper, we propose DialectMoE, a Chinese multi-dialects speech recognition model based on Mixture-of-Experts (MoE) in a low-resource conditions. Specifically, DialectMoE assigns input sequences to a set of experts using a dynamic routing algorithm, with each expert potentially trained for a specific dialect. Subsequently, the outputs of these experts are combined to derive the final output. Due to the similarities among dialects, distinct experts may offer assistance in recognizing other dialects as well. Experimental results on the Datatang dialect public dataset show that, compared with the baseline model, DialectMoE reduces Character Error Rate (CER) for Sichuan, Yunnan, Hubei and Henan dialects by 23.6%, 32.6%, 39.2% and 35.09% respectively. The proposed DialectMoE model demonstrates outstanding performance in multi-dialects speech recognition.

## 1 Introduction

The application domains of speech recognition technology are extensive, encompassing diverse fields such as voice assistants, smart homes, and automotive voice interaction, among others. Thanks to the advancements in deep learning, Automatic Speech Recognition (ASR) systems have made remarkable strides in recognizing Mandarin speech (Malik et al., 2021; Wang et al., 2019; Alharbi et al., 2021).

Dialect serves as a prevalent mode of everyday communication among the Chinese populace. However, the performance of ASR systems remains limited in dialect speech, posing a significant challenge in the field of speech recognition technology (Hinsvark et al., 2021; Alsharhan et al., 2020) due to the inherent variations and distinct characteristics in pronunciation among dialects and Mandarin. Therefore, improving the accuracy and adaptability of Chinese ASR systems is significant and meaningful for multi-dialect. Our study mainly focuses on Chinese dialects, the proposed method can also be generalized to other dialects.

Chinese dialects are typically classified into ten main categories, each exhibiting notable differences in pronunciation, tone, vocabulary, and grammar (Ho et al., 2015). Chinese is a tonal language, where each character corresponds to a specific tone, a feature that is prevalent in most of its dialects as well. The pronunciation of a given Chinese character with different tones imparts markedly distinct meanings. This underscores the profound significance of tones in the comprehension of Chinese phonetics. (Ho et al., 2015; Sproat et al., 2004). Figure 1 depicts the tonal distinctions among Standard Mandarin,

*Shengxiang Gao (Corresponding Author): gaoshengxiang.yn@foxmail.com
This research is funded in part by the National Natural Science Foundation of China (No.62376111, U23A20388, U21B2027, 62366027), Key Research and Development Program of Yunnan Province (202303AP140008, 202302AD080003, 202401BC070021, 202103AA080015), Science and Technology Talent and Platform Program of Yunnan Province (202105AC160018).
©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1148-1159, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers

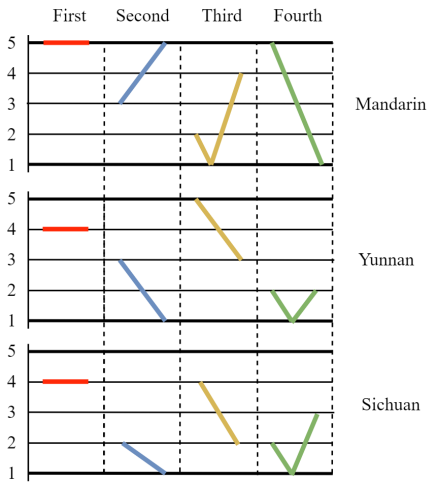



Figure 1: The tonal distinctions among Standard Mandarin, Yunnan dialect, and Sichuan dialect.

Yunnan dialect, and Sichuan dialect. It is evident that notable in tone between Standard Mandarin and Yunnan dialect as well as Sichuan dialect in the second, third, and fourth tones. However, Yunnan dialect and Sichuan dialect exhibit a pronunciation similarity in specific tones. The change in tone reveals differences and similarities between Standard Mandarin and various dialects. Hence, considering both the differences and similarities in pronunciation among various dialects alongside Standard Mandarin becomes crucial for the advancement of Chinese speech recognition systems.

In recent years, numerous researches have focused on tackling the challenge of poor performance in dialect speech recognition models (Li et al., 2018; Ren et al., 2019; Zhang et al., 2022). The traditional way is based on different modeling methods to improve the effect of dialect speech recognition. Humphries (1996) employed an adaptive method that utilizes a pronunciation vocabulary with dialect data to capture differences between standard and dialect pronunciations. Li (2019) proposed a novel method for modeling Chinese characters based on radicals, effectively addressing the issue of dialect modeling difficulty. This method significantly reduces the required size of radical dictionaries compared to ordinary character dictionaries. Recently, multitask-based methods have been widely used in the task of dialect speech recognition. Compared with the traditional method, the multi-task learning method is more efficient. Elfeky (2016) proposed constructing a dialect classification model and a separate speech recognition model for each dialect. The dialect classification model is used to select the corresponding dialect speech recognition model. Dan (2022) proposed a multi-task training strategy that combines dialect classification with dialect speech recognition, bridging the substantial gap between Mandarin and dialect acoustic properties. However, these investigations are contingent upon extensive dialectal datasets and do not examine the potential influence of commonalities among various dialects on model performance.

To construct a reliable dialect speech recognition model in low-resource conditions, Jiang (2023) introduced a transfer learning-based approach, it involves a model trained on Mandarin and fine-tunes with small-scale dialect data. However, relying solely on transfer learning may not adequately capture the distinctions between dialects and Mandarin. Wang (2023) proposed the Aformer model with multi-stage training strategy, which can capture diverse acoustic information in different training stage, enabling the model to effectively adapt to dialect data. The aforementioned studies focus on the training strategy and model expansion, they do not fully consider the differences and similarities between dialects and Mandarin.

In this paper, we present DialectMoE, a multi-dialect speech recognition model based on Mixture-of-Experts (MoE), aimed at improving the performance of multi-dialects speech recognition in low-resource conditions. DialectMoE is architecturally structured with dual encoders: a dialect encoder and a general encoder. The main contributions of the paper include:

- We propose a three-stage training methodology designed to enhance the model's adaptability in

addressing low-resource multi-dialect scenarios through different stages. Detailed specifics will be expounded upon in Section 3.3.

- We introduce MoE layers and enhance the dynamic routing algorithm to enable the combination of acoustic features from both the input sequence and the dialect encoder during the expert selection process.
- The experiment results show that DialectMoE reduces Character Error Rate (CER) compared to the baseline model for Sichuan, Yunnan, Hubei and Henan dialects by 23.6%, 32.6%, 39.2% and 35.09%, respectively.

## 2 Related Work

### 2.1 Conformer-based ASR

The Conformer, a convolution-augmented Transformer introduced in (Gulati et al., 2020), has been widely acknowledged as the state-of-the-art end-to-end ASR technology owing to its exceptional performance in ASR tasks. In recent years, several researchers have proposed Conformer variants (Peng et al., 2022; Sehoon et al., 2023) to further enhance the capabilities of speech recognition. The Conformer module comprises two feed-forward modules, a multi-head self-attention module, and a convolution module. The output $y$ of one Conformer block for a given input $x$ can be defined as follows:

$$\hat{x} = x + \frac{1}{2}\,\mathrm{FFN}_1(x) \tag{1}$$

$$\tilde{x} = \hat{x} + \mathrm{MHSA}(\hat{x}) \tag{2}$$

$$\bar{x} = \tilde{x} + \mathrm{Conv}(\tilde{x}) \tag{3}$$

$$y = \mathrm{LN}(\bar{x} + \frac{1}{2}\,\mathrm{FNN}_2(\bar{x})) \tag{4}$$

where $\mathrm{FNN}_1$ denotes the first feedforward module, $\mathrm{FNN}_2$ denotes the second feedforward network, $\mathrm{MHSA}$ denotes the multi-head self-attention module, $\mathrm{Conv}$ denotes the convolution module, and $\mathrm{LN}$ denotes layer normalization. For additional information regarding the Conformer ASR model, please refer to (Gulati et al., 2020).

During Conformer training, the Joint CTC-Attention loss function (Hori et al., 2017) is utilized. This loss function is commonly used in present-day speech recognition technology. In this paper, the joint CTC-Attention loss is incorporated into the total loss function. The loss function is outlined as follows:

$$\mathcal{L}_{all} = (1-\lambda)\mathcal{L}_{att} + \lambda\mathcal{L}_{ctc} \tag{5}$$

where $\mathcal{L}_{att}$ denotes the decoding loss of the Attention decoder, and $\mathcal{L}_{ctc}$ denotes the CTC loss, $\lambda$ is a hyper parameter which denotes the weight of these two loss function.

### 2.2 Mixture-of-Experts Based Speech Recognition

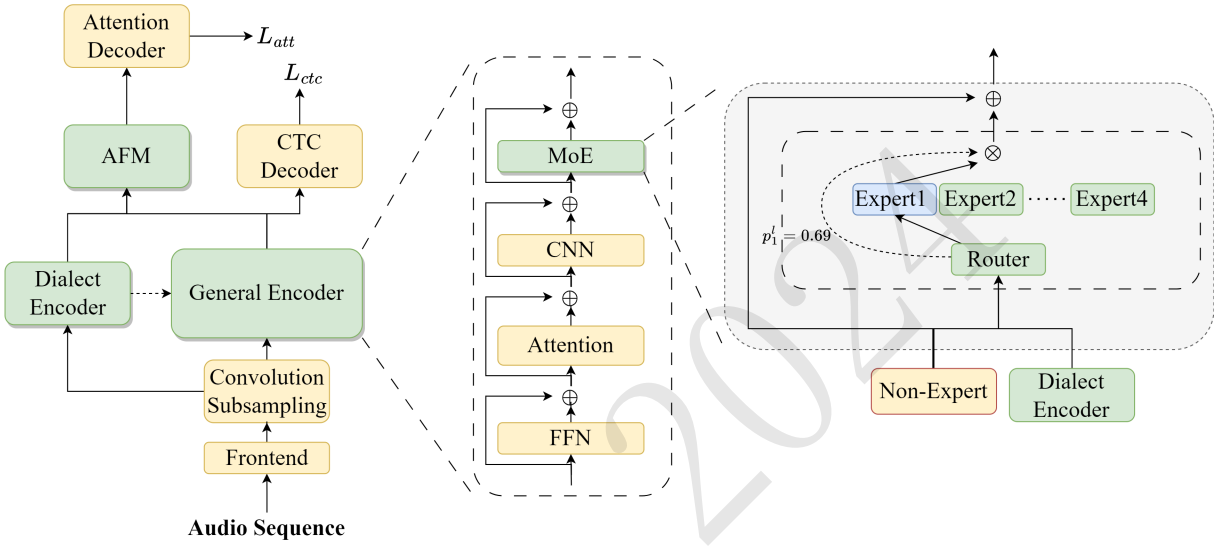
The MoE based methods offer a solution for more efficient training and inference by selectively activating different experts in the model based on different inputs (Jacobs et al., 1991; Shazeer et al., 2017). This enables the model to adapt to a wide range of inputs and scale to more parameters while maintaining a consistent computational cost. The MoE based models have demonstrated their effectiveness in natural language processing (Fedus et al., 2022; Du et al., 2022) and computer vision (Riquelme et al., 2021; Fan et al., 2022).

In real-world applications, speech recognition systems are required to adeptly cope with diverse input conditions, encompassing variations in speakers, accents, and acoustic environments (Zilvan et al., 2021). However, conventional speech recognition models have a fixed computational cost and cannot adapt to the complexity of input instances. You (2021; 2022) explore the MoE based model for speech recognition, named SpeechMoE, and propose a new router architecture which integrates additional global domain and embeddings into router input to promote adaptability. Additionally, a multilingual speech recognition network (MoLE) was introduced (Kwon et al., 2023) to analyze audio input

CCL 2024

data from multiple languages and identify expert networks suitable for each language. Simultaneously, a language-independent expert network was also introduced, and the selected expert network and the language-independent expert network collectively fulfill the language requirements necessary for effective speech recognition.

Employing the MoE mechanism to determine expert activation during the forward propagation process manifests a notable capacity for accommodating the inherent variability in multi-dialectal speech across different input sequences. Nevertheless, the conventional MoE paradigm relies solely on the input sequence for expert selection, and the information in the present input does not inherently ensure the optimal suitability of the selected experts. Therefore, the incorporation of supplementary dialectal information to facilitate expert selection stands forth as a judicious resolution, enhancing the precision and adaptability of the chosen experts to the distinctive intricacies characterizing the multi-dialectal speech context. Furthermore, the exploration of MOE-based methods in the domain of multi-dialect speech recognition remains limited.




(a) DialectMoE (b) DialectMoE block

Figure 2: (a) DialectMoE overall architecture, where the general encoder consists of N DialcetMoE blocks. (b) Architecture of the DialectMoE encoder block module.

## 3 DialectMoE

### 3.1 Overall Architecture of DialectMoE

The overall architecture of DialectMoE is shown in Figure 2(a). The original audio sequence undergoes preprocessing by the frontend module to extract filter bank (FBank) features (Singh et al., 2020). Subsequently, the convolutional downsampling is applied to temporally downsample the audio feature sequence. The dialect encoder, consisting of 6 layers of vanilla Conformer encoder, captures dialect information from the feature sequences. The general encoder, which comprises 12 layers of DialectMoE encoder blocks, is responsible for capturing speech information in a dialect-agnostic manner. Both encoders share the same input but focus on different aspects of information.

The detailed structure of the DialectMoE block is presented in Figure 2(b). With the DialectMoE block, the input sequence is first passed through the Feed-Forward Network (FFN) layer, followed by Attention and Convolutional Neural Network (CNN) layer to extract global and local information, respectively. Then the appropriate expert within the MoE layer is selected based on the dynamic routing. The output of experts are multiplied by the weight assigned by the router layer.

Compared to the widely used vanilla Conformer block (Gulati et al., 2020), our DialectMoE block incorporates MoE layers to address complex and variable scenarios encountered in real-world situations.

CCL 2024

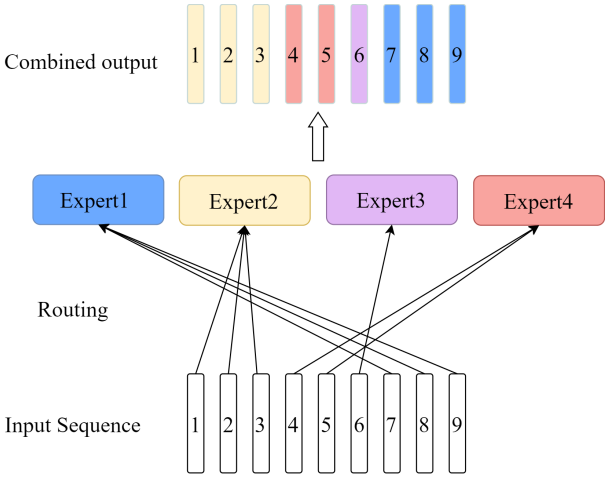



Figure 3: Illustration of dynamic routing algorithm.

The dialect information captured by the dialect encoder is weighted by the router layer, which enables the router layer to choose more appropriate experts based on both dialect features and general features obtained from two encoders. This dynamic routing mechanism proves more effective in intricate speech scenarios, especially those involving multiple dialects.

### 3.2 Dialect Adaptive Dynamic Expert Routing

In the context of multi-dialect speech recognition, effectively addressing the diversity of dialectal variations is crucial. We present a novel dynamic routing algorithm aimed at enhancing the adaptability and generalization of the model to diverse dialects. The proposed algorithm leverages the input sequences from the current MoE layer and the dialect information provided by the dialect encoder to select appropriate experts. To evaluate the impact of different dialect embedding on routing, we explore the following three strategies: utilizing the embedding (**embed**) independently, and both concatenating (**concat**) and adding (**add**) the embeddings to the output of the convolutional layers. The output of the dialect encoder is denoted as $\mathbf{X}^{D}_{encoder} \subseteq \mathbb{R}^{T\times d}$, where $T$ represents the sequence length and $d$ denotes the feature dimension. Assuming that there are $N$ experts, the output $r \subseteq \mathbb{R}^{T\times N}$ of the routing layer can be defined as follows:

$$r = W_r \cdot \mathrm{Concat}(\bar{x}, \mathbf{X}^{D}_{encoder}) \tag{6}$$

$$r = W_r \cdot \mathrm{Add}(\bar{x}, \mathbf{X}^{D}_{encoder}) \tag{7}$$

$$r = W_r \cdot \mathbf{X}^{D}_{encoder} \tag{8}$$

where $W_r$ represents the weight parameter of the router layer, and $\bar{x}$ denotes the output of the convolution module. These three dynamic routing strategies are the ones we consider employing. It is worth noting that while the general router layer selects experts based on the input sequence $\bar{x}$, the algorithm we designed intuitively makes more sense as it incorporates the output of the dialect encoder to select the most suitable expert.

The router layer selects the expert with the highest probability through dynamic routing, which is based on the router output $r$. The dynamic routing probability is then defined as follows:

$$p_i = \frac{exp^{r_i}}{\sum_{j=1}^{N} exp^{r_j}} \tag{9}$$

where $p_i \subseteq \mathbb{R}^{T\times N}$ is the probability that the $i$ expert is selected, the output $\mathbf{O}_{moe} \subseteq \mathbb{R}^{T\times d}$ of the MoE layer can be formally defined as follows:

$$\mathbf{O}_{moe} = p_i \cdot E_i(\bar{x}) \tag{10}$$

CCL 2024

where $E_i$ is the output of the $i$ expert selected. Figure 3 illustrates the process of dynamic routing.

In order to incorporate dialect information into the decoder, DialectMoE incorporates an information fusion step by combining the outputs of two separate encoders. This fusion process, illustrated as the Acoustic Fusion Module (**AFM**) in Figure 2(a), occurs prior to transmitting the results to the decoder. The fusion process is defined as follows:

$$\mathbf{X}^{A}_{encoder} = \text{Concat}(\mathbf{X}^{G}_{encoder}, \mathbf{X}^{D}_{encoder}) \tag{11}$$

where $\mathbf{X}^{A}_{encoder}$ denotes the result of fusion of information output by two two different encoders, and $\mathbf{X}^{G}_{encoder}$ denotes the result output by a general encoder.

The comprehensive loss function for speech recognition comprises the combined CTC-Attention loss (Hori et al., 2017), as explained previously, along with the supplementary balance loss (Fedus et al., 2022). The complete formulation of the loss function is as follows:

$$\mathcal{L}_{all} = \lambda\mathcal{L}_{ctc} + (1-\lambda)\mathcal{L}_{att} + \alpha\mathcal{L}_{b} \tag{12}$$

where $\alpha$ is the weight of the balance loss ($\alpha = 0.1$) and $\lambda$ is the weight of the speech recognition loss ($\lambda = 0.3$), $\mathcal{L}_b$ denotes the balance loss.

### 3.3 Training Strategies

Considering the significant disparity in the quantities of Mandarin and dialect data, low-resource dialect speech recognition scenarios commonly exhibit limited labeled dialect speech data, typically ranging from a few to tens of hours. This insufficiency hampers the development of a reliable speech recognition model. To address this issue, this study introduces a multi-stage training strategy. The training process encompasses the following sequential steps:

1. Pre-training: The Conformer model is used as a general encoder for DialectMoE to implement pre-training on Mandarin datasets. The pre-training step allows the model to capture various common speech features, thus reducing the complexity of learning for the dialect recognition task.

2. Training Dialect Encoder: A Conformer Encoder is initialized as a dialect encoder and is trained on the dialect classification task using both dialect and Mandarin data. The objective of this step is to enable the dialect encoder to learn the acoustic differences between multiple dialects, assisting the general encoder in dialect speech recognition tasks.

3. Training DialectMoE: The parameters of the dialect encoder are frozen, and the second feedforward network layer in the pre-trained Conformer model is initialized with $N$ experts. Use only low-resource dialect training data to train the final DialectMoE model.

By pre-training, the initial model acquires a substantial set of effective parameters, thereby conferring notable advantages for last training stages. In the second phase, the dialect encoder is trained on a dialect identification task, enabling it to focus on differences between multiple dialects and Mandarin. In the last stage, only multi-dialect data is used for training. This approach enhances DialectMoE's capability to adeptly capture shared acoustic characteristics across multiple dialects. The proposed method is evaluated based on extensive comparison and ablation experiments, which are comprehensively detailed in Section 4.

## 4 Experiments

### 4.1 Datasets

The AIShell-1 dataset (Bu et al., 2017) serves as the Mandarin speech corpus in this study. This Mandarin speech dataset is widely employed in the field of Chinese speech recognition technology.

For the Chinese dialect dataset, an open-source dataset provided by Datatang[1] is utilized in this study. It comprises a training set of 30 hours of Sichuan and Yunnan dialects and a test set of 1.5 hours featuring

[1] https://www.datatang.com

| Dataset | Train(h) | Test(h) |
|---|---|---|
| Aishell | 164 | 10 |
| Sichuan (**SC**) | 28.5 | 1.5 |
| Yunnan (**YN**) | 28.5 | 1.5 |
| Henan (**HN**) | 0 | 1.5 |
| Hubei (**HB**) | 0 | 1.5 |

Table 1: Details of both Dialect and Mandarin datasets.

Henan and Hubei dialects. Within this study, Sichuan (**SC**) and Yunnan (**YN**) dialects are used to test the adaptability of the model to multi-dialect data, and Henan (**HN**) and Hubei (**HB**) dialects are used to test the generalization of the model to multi-dialect data. More details are shown in Table 1.

### 4.2 Experiment Setup

All experiments were conducted using the Wenet (Zhang et al., 2017) end-to-end speech toolkit. Our methodology involved extracting an 80-dimensional log-Mel filter bank (FBank) as the acoustic input feature, with a window size of 25 ms and a step size of 10 ms. To ensure feature normalization, we applied mean and variance normalization (MVN) calculated from the training set on Fbank. To augment the low-resource dialect data, we employed speed perturbation and SpecAugment (Park et al., 2017) techniques. No additional language models were incorporated into the experiments.

For the pre-training model, we utilized a Conformer encoder trained on the Mandarin dataset. The general encoder of DialectMoE consists of 12 Conformer encoder layers with a feed-forward dimension of 2048 and an attention dimension of 256, employing 4 self-attention heads. This model was trained using the Adam optimizer (Kingma et al., 2014). Furthermore, we adopted the warmup learning schedule (Gotmare et al., 2018) for the initial 25K training iterations, initializing the learning rate at 0.002, and set the label smoothing (Szegedy et al., 2016) weight and dropout to 0.1 for model regularization. The decoder consists of a 6-layer Transformer, while the dialect encoder comprises a 6-layer Conformer encoder. The cross-entropy loss for classifying dialect is always applied.

The proposed DialectMoE is initialized with the pre-trained general encoder, dialect encoder, and decoder. The second feedforward layer in each Conformer layer of the general encoder is initialized as an $N$ expert ($N = 4$), with the expert parameters being the pre-trained feedforward network parameters. Training employed the same Adam optimizer, and the number of warm-up steps in the pre-training learning plan was adjusted to 10000, with an initial learning rate of 0.001.

### 4.3 Main Results

| ID | Model | Params(M) | SC | YN | HN | HB |
|---|---|---|---|---|---|---|
| M1 | Conformer (Mandarin) | 46.1 M | 82.75 | 81.42 | 82.26 | 87.47 |
| M2 | FT-Conformer | 46.1 M | 15.60 | 14.06 | 54.91 | 57.18 |
| M3 | Conformer (Mandarin+Dialect) | 46.1 M | 13.86 | 12.02 | 41.28 | 48.37 |
| M4 | MT-Conformer(DID+ASR) | 47.2 M | 17.79 | 15.64 | 47.78 | 56.55 |
| M5 | MT-Conformer(DID&ASR) | 46.2 M | 16.09 | 14.05 | 41.28 | 48.37 |
| M6 | Aformer | 68.3 M | 13.21 | 12.76 | 35.32 | 39.89 |
| M7 | DialectMoE | 93.8 M | **11.91** | **9.84** | **33.38** | **37.11** |

Table 2: CER(%) on Chinese Dialect ASR task. **FT** represents the fine-tuning step and **MT** represents the multitask-based approach.

In this paper, we meticulously design comparison experiments with other speech recognition models to showcase the effectiveness of our proposed method. The experimental results presented in this study were reproduced using the open-source speech processing toolkit Wenet (Zhang et al., 2017). Table 2 illustrates the performance of each ASR model in dialect speech recognition under low-resource

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1148-1159, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

conditions. The evaluation metric employed is the Character Error Rate (CER).

**M1** represents the Conformer model that was exclusively pre-trained on the Aishell Mandarin dataset, consisting of 178 hours of data.

**M2** denotes the model fine-tuned from M1 using the low-resource dialect dataset.

**M3** corresponds to the model trained directly on the combined dataset of both dialect and Mandarin speech.

**M4** and **M5** refer to the multi-tasking models trained on the combined dataset, with a distinction that M4 predicts the dialect category in the encoder while the decoder focuses on recognizing the speech text, whereas M5 predicts both the dialect category and the speech text in the decoder.

**M6** represents the multi-pass model proposed in (Wang et al., 2023) for the training of the Aformer.

**M7** signifies the DialectMoE model proposed in this paper.

The results obtained from the M1 model demonstrate a notably poor performance in recognizing dialectal speech within the Mandarin speech recognition model. However, by fine-tuning the M1 model with dialect data, the CER of the M2 model for Sichuan and Yunnan dialects is significantly reduced, although further optimization is still required. To address this, our paper proposes the DialectMoE model, which surpasses existing studies and baselines in terms of performance. In comparison to the fine-tuned model of M2, the DialectMoE model exhibits a reduction in CER of 23.6% and 32.6% for the Sichuan and Yunnan dialects, respectively. Additionally, it achieves a reduction of 39.2% and 35.09% for the Henan and Hubei dialects, respectively.

### 4.4 Ablation Studies

#### 4.4.1 Ablation of dynamic routing strategy

| Strategy | SC | YN | HN | HB |
|---|---|---|---|---|
| normal | 13.43 | 11.83 | 35.22 | 38.67 |
| embed | 12.53 | 10.20 | 34.95 | 38.65 |
| concat | 12.18 | 9.97 | 33.55 | **37.09** |
| add | 12.41 | 10.36 | 34.19 | 37.23 |
| normal+fusion | 13.92 | 12.19 | 35.27 | 38.72 |
| embed+fusion | 12.23 | 10.21 | 33.50 | 37.55 |
| concat+fusion | **11.91** | **9.84** | **33.38** | 37.11 |
| add+fusion | 12.49 | 10.13 | 33.66 | 37.65 |

Table 3: Ablation of different routing strategy.

This paper incorporates ablation experiments to investigate the effectiveness of the proposed dynamic routing algorithm and model design. Table 3 presents the impact of utilizing different dynamic routing algorithms and the merging of two encoder outputs before the decoder. In ”normal”, the dynamic routing algorithm proposed in this paper is not employed, and the experts are directly selected based on the input sequence, similar to the approach in (Fedus et al., 2022). The strategy column in Table 3 indicates the usage of different dynamic routing algorithms:“**embed**” signifies the utilization of only the dialect encoder outputs, “**concat**” denotes the concatenation of the dialect encoder outputs with the input sequence, and “**add**” indicates the summation of the dialect encoder outputs with the input sequence. The “**fusion**” entry indicates whether or not the two encoder outputs should be fused before reaching the decoder., whether they go through AFM. The experiments employing the “**concat+fusion**” strategy along with the fusion of the two encoder outputs demonstrate optimal results across the four different dialect test sets.

#### 4.4.2 Ablation of experts number

To investigate the impact of initializing a different number of experts in DialectMoE on the overall model performance, we conducted an experiment with varying numbers of experts, specifically 2, 4, and 8. The experimental results, as shown in Table 4, highlight that the model size increases with an increasing number of experts. However, the common notion that a larger number of model participants leads to

CCL 2024

| Model | Params(M) | SC | YN | HN | HB |
|---|---|---|---|---|---|
| MoE-2e | 68.5M | 12.39 | 10.21 | 33.84 | 38.37 |
| MoE-4e | 93.8M | **11.91** | **9.84** | **33.38** | **37.11** |
| MoE-8e | 134.5M | 12.94 | 10.44 | 34.09 | 38.26 |

Table 4: Ablation of experts number.

improved performance does not hold true under low-resource conditions. The results indicate that, for low-resource dialect data, an excessive number of experts does not enhance model performance; in fact, it diminishes it. Experimental evidence supports the conclusion that setting the number of experts to 4 is more appropriate in this context. It is noteworthy that when the number of experts is set to 2, the number of model parameters matches the number of Aformer (Wang et al., 2023) parameters. However, despite this similarity, our results outperform the baseline. This finding further validates the efficacy and correctness of our proposed method.

#### 4.4.3 Ablation of the number of experts selected

| Top-k | Time(s) | SC | YN | HN | HB |
|---|---|---|---|---|---|
| 4 | 1.46s | 12.01 | 10.03 | **33.31** | **37.08** |
| 2 | 0.94s | 11.93 | **9.81** | 33.57 | 37.24 |
| 1 | **0.68s** | **11.91** | 9.84 | 33.38 | 37.11 |

Table 5: Ablation of the number of experts selected.

In the vanilla MoE, a top-k approach is employed to select a combination of k experts for routing the input sequence. However, in this paper, a $\mathrm{Softmax}$ approach, specifically top-1, is utilized. To further investigate the effectiveness of the proposed dynamic routing algorithm, experiments were conducted to explore the impact of the number of selected experts. As presented in Table 5, when the number of selected experts is set to 4, there is an improvement in performance for dialects that are not part of the training dataset (Henan and Hubei dialects). This suggests that increasing the number of selected experts can enhance the model's generalization to external data. The model's performance remains similar when the number of selected experts is 2 or 1. However, it is worth noting that the decoding time for a single speech increases by approximately 53% when the number of selected experts is 4 compared to when it is 1. This indicates that the number of selected experts has a minimal impact on the model's performance but significantly affects decoding efficiency, which is crucial for a robust speech recognition system.

### 4.5 Layer-Wise Analysis of Experts

In Figure 4, We randomly extracted 100 samples from the test sets of Sichuan and Kunming dialects to visualize the expert weights. We can observe certain patterns in the weights. Across the initial three layers of the model, both dialects manifest a heightened degree of distinctiveness in expert selection, indicative of specific groups of experts concentrating exclusively on dialect-specific information. Within the intermediate layers of the model, expert weights display a diminished prominence, yet discernible differences persist in the expert weights associated with the two dialects. This observation suggests that varying combinations of experts implicitly encapsulate distinctive information pertaining to dialectal variations. In the concluding three layers of the model, the deployed experts exhibit near-identical characteristics, thereby indirectly affirming the model's proficiency in capturing shared features between Sichuan and Kunming dialects.

## 5 Conclusion

In this manuscript, we present a multi-dialectal speech recognition model based on MoE termed Dialect-MoE. Structurally, it incorporates a dual-encoder architecture, wherein the general encoder is dedicated to acquiring general acoustic representations, and the dialect encoder is specialized for acquiring acoustic representations across various dialects. A refinement in the dynamic routing strategy within the MoE

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1148-1159, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

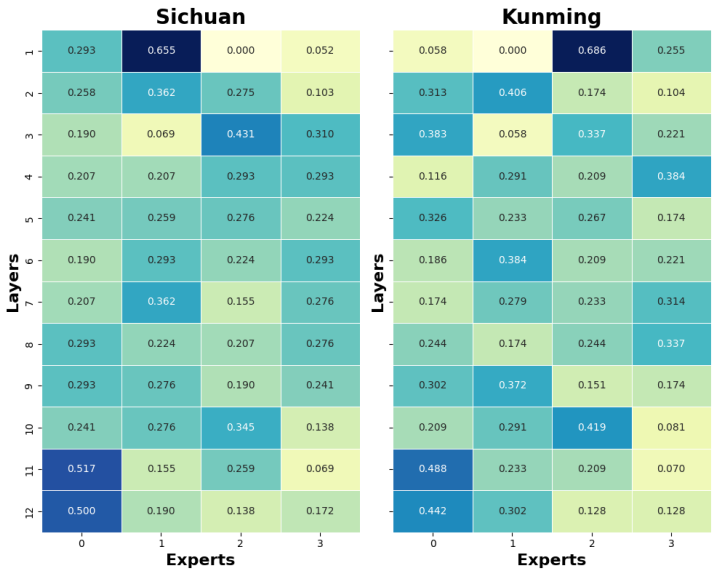



Figure 4: The expert weights are visualized on Sichuan dialect and Kunming dialect.

layer of the universal encoder has been introduced to enable the selection of appropriate experts based on the acoustic information specific to the dialect in the input sequence. Furthermore, we propose a three-stage training methodology to facilitate DialectMoE in learning distinct tasks at different phases, thereby enhancing its adaptability and performance across varying aspects of the multi-dialectal speech recognition task. Experimental results demonstrate that the proposed DialectMoE model achieves remarkable performance in multi-dialects speech recognition tasks.

## 6 Limitations

While the MoE-based approach can effectively enhance model performance, it inherently results in an increase in the number of model parameters. This increase in parameters can lead to higher training costs and occupies more space on, e.g., a GPU, which are inevitable consequences. Therefore, it is imperative to conduct further research on model compression techniques to mitigate these issues.

## References

Alharbi, Sadeen and Alrazgan, Muna and Alrashed, Alanoud and Alnomasi, Turkiayh and Almojel, Raghad and Alharbi, Rimah and Alharbi, Saja and Alturki, Sahar and Alshehri, Fatimah and Almojil, Maha. 2021. Automatic speech recognition: Systematic literature review. *IEEE Access*,9:131858–131876.

Alsharhan, Eiman and Ramsay, Allan. 2020. Investigating the effects of gender, dialect, and training size on the performance of Arabic speech recognition. *Language Resources and Evaluation*,54:975–998.

Bu, Hui and Du, Jiayu and Na, Xingyu and Wu, Bengu and Zheng, Hao. 2017. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. *2017 20th conference of the oriental chapter of the international coordinating committee on speech databases and speech I/O systems and assessment (O-COCOSDA)*,pages 1–5.

Dan, Zhengjia and Zhao, Yue and Bi, Xiaojun and Wu, Licheng and Ji, Qiang. 2022. Multi-task transformer with adaptive cross-entropy loss for multi-dialect speech recognition. *Entropy*,24(10):1429.

Du, Nan and Huang, Yanping and Dai, Andrew M and Tong, Simon and Lepikhin, Dmitry and Xu, Yuanzhong and Krikun, Maxim and Zhou, Yanqi and Yu, Adams Wei and Firat, Orhan and others. 2022. Glam: Efficient scaling of language models with mixture-of-experts. *International Conference on Machine Learning*,pages 5547–5569.

Elfeky, Mohamed and Bastani, Meysam and Velez, Xavier and Moreno, Pedro and Waters, Austin. 2016. Towards acoustic model unification across dialects. *2016 IEEE Spoken Language Technology Workshop (SLT)*,pages 624–628.

Fan, Zhiwen and Sarkar, Rishov and Jiang, Ziyu and Chen, Tianlong and Zou, Kai and Cheng, Yu and Hao, Cong and Wang, Zhangyang and others. 2022. $M^3$vit: Mixture-of-experts vision transformer for efficient multi-task learning with model-accelerator co-design. *Advances in Neural Information Processing Systems*,35:28441–28457.

Fedus, William and Zoph, Barret and Shazeer, Noam. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *The Journal of Machine Learning Research*,23(1):5232–5270.

Gotmare, Akhilesh and Keskar, Nitish Shirish and Xiong, Caiming and Socher, Richard. 2018. A closer look at deep learning heuristics: Learning rate restarts, warmup and distillation. *arXiv preprint arXiv:1810.13243*.

Gulati, Anmol and Qin, James and Chiu, Chung-Cheng and Parmar, Niki and Zhang, Yu and Yu, Jiahui and Han, Wei and Wang, Shibo and Zhang, Zhengdong and Wu, Yonghui and others. 2020. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.

Hinsvark, Arthur and Delworth, Natalie and Del Rio, Miguel and McNamara, Quinten and Dong, Joshua and Westerman, Ryan and Huang, Michelle and Palakapilly, Joseph and Drexler, Jennifer and Pirkin, Ilya and others. 2021. Accented speech recognition: A survey. *arXiv preprint arXiv:2104.10747*.

Ho, Dah-an. 2015. Chinese dialects. *The Oxford handbook of Chinese linguistics*,pages 149–159.

Hori, Takaaki and Watanabe, Shinji and Hershey, John R. 2017. Joint CTC/attention decoding for end-to-end speech recognition. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*,pages 518–529.

Humphries, Jason J and Woodland, Philip C and Pearce, D. 1996. Using accent-specific pronunciation modelling for robust speech recognition. *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*,4:2324–2327.

Jacobs, Robert A and Jordan, Michael I and Nowlan, Steven J and Hinton, Geoffrey E. 1991. Adaptive mixtures of local experts. *Neural computation*,3(1):79–87.

Jiang Rui. 2023. Chinese Dialect Recognition Based on Transfer Learning. *INTERSPEECH*.

Kingma, Diederik P and Ba, Jimmy. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kwon, Yoohwan and Chung, Soo-Whan. 2023. MoLE: Mixture Of Language Experts For Multi-Lingual Automatic Speech Recognition. *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*,pages 1–5.

Li, Bo and Sainath, Tara N and Sim, Khe Chai and Bacchiani, Michiel and Weinstein, Eugene and Nguyen, Patrick and Chen, Zhifeng and Wu, Yanghui and Rao, Kanishka. 2018. Multi-dialect speech recognition with a single sequence-to-sequence model. *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*,pages 4749–4753.

Li, Sheng and Lu, Xugang and Ding, Chenchen and Shen, Peng and Kawahara, Tatsuya and Kawai, Hisashi. 2019. Investigating Radical-Based End-to-End Speech Recognition Systems for Chinese Dialects and Japanese. *INTERSPEECH*,pages 2200–2204.

Malik, Mishaim and Malik, Muhammad Kamran and Mehmood, Khawar and Makhdoom, Imran. 2021. Automatic speech recognition: a survey. *Multimedia Tools and Applications*,80:9411-9457.

Park, Daniel S and Chan, William and Zhang, Yu and Chiu, Chung-Cheng and Zoph, Barret and Cubuk, Ekin D and Le, Quoc V. 2017. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*.

Peng, Yifan and Dalmia, Siddharth and Lane, Ian and Watanabe, Shinji. 2022. Branchformer: Parallel mlp-attention architectures to capture local and global context for speech recognition and understanding. *International Conference on Machine Learning*,pages 17627–17643.

Ren, Zongze and Yang, Guofu and Xu, Shugong. 2019. Two-stage training for chinese dialect recognition. *arXiv preprint arXiv:1908.02284*.

Riquelme, Carlos and Puigcerver, Joan and Mustafa, Basil and Neumann, Maxim and Jenatton, Rodolphe and Susano Pinto, André and Keysers, Daniel and Houlsby, Neil. 2021. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*,34:8583–8595.

Sehoon and Gholami, Kim, Amir and Shaw, Albert and Lee, Nicholas and Mangalam, Karttikeya and Malik, Jitendra and Mahoney, Michael W and Keutzer, Kurt. 2023. Squeezeformer: An efficient transformer for automatic speech recognition. *Advances in Neural Information Processing Systems*,35:9361-9373.

Shazeer, Noam and Mirhoseini, Azalia and Maziarz, Krzysztof and Davis, Andy and Le, Quoc and Hinton, Geoffrey and Dean, Jeff. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Singh, Yuvika and Pillay, Anban and Jembere, Edgar. 2020. Features of speech audio for accent recognition. *2020 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD)*,pages 1–6.

Sproat, Richard and Gu, Liang and Li, Jing and Zheng, Yanli and Su, Yi and Zhou, Haolang and Bramsen, Philip and Kirsch, MIT David and Shafran, Izhak and Tsakalidis, Stavros and others. 2004. Dialectal Chinese speech recognition. *CLSP Summer Workshop*.

Szegedy, Christian and Vanhoucke, Vincent and Ioffe, Sergey and Shlens, Jon and Wojna, Zbigniew. 2016. Rethinking the inception architecture for computer vision. *Proceedings of the IEEE conference on computer vision and pattern recognition*,pages 2818–2826.

Wang, Dong and Wang, Xiaodong and Lv, Shaohe. 2019. An overview of end-to-end automatic speech recognition. *Symmetry*,11(8):1018.

Wang, Xuefei and Long, Yanhua and Li, Yijie and Wei, Haoran. 2023. Multi-pass Training and Cross-information Fusion for Low-resource End-to-end Accented Speech Recognition. *INTERSPEECH*.

You, Zhao and Feng, Shulin and Su, Dan and Yu, Dong. 2021. Speechmoe: Scaling to large acoustic models with dynamic routing mixture of experts. *INTERSPEECH*.

You, Zhao and Feng, Shulin and Su, Dan and Yu, Dong. 2022. Speechmoe2: Mixture-of-experts model with improved routing. *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*,pages 7217–7221.

Zahner, Katharina and Kutscheid, Sophie and Braun, Bettina. 2019. Alignment of f0 peak in different pitch accent types affects perception of metrical stress. *Journal of Phonetics*,74:75–95.

Zhang, Binbin and Wu, Di and Peng, Zhendong and Song, Xingchen and Yao, Zhuoyuan and Lv, Hang and Xie, Lei and Yang, Chao and Pan, Fuping and Niu, Jianwei. 2017. Wenet 2.0: More productive end-to-end speech recognition toolkit. *arXiv preprint arXiv:2203.15455*.

Zhang, Fengrun and Xie, Xiang and Quan, Xinyue. 2022. Chinese Dialect Speech Recognition Based on End-to-end Machine Learning. *2022 International Conference on Machine Learning, Control, and Robotics (MLCR)*,pages 14–18.

Zilvan, Vicky and Heryana, Ana and Yuliani, Asri Rizki and Krisnandi, Dikdik and Yuwana, R Sandra and Pardede, Hilman F. 2021. Front-end Based Robust Speech Recognition Methods: A Review. *Proceedings of the 2021 International Conference on Computer, Control, Informatics and Its Applications*,pages 136–140.

CCL 2024

# Distinguishing Neural Speech Synthesis Models Through Fingerprints in Speech Waveforms

**Chu Yuan Zhang[1,2], Jiangyan Yi[1], Jianhua Tao[3,4], Chenglong Wang[1], Xinrui Yan[1,2]**

[1] Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2] University of Chinese Academy of Sciences, Beijing, China
[3] Department of Automation, Tsinghua University, Beijing, China
[4] Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing, China
{zhangchuyuan2021, yanxinrui2021}@ia.ac.cn,
{jiangyan.yi, chenglong.wang}@nlpr.ia.ac.cn,
jhtao@tsinghua.edu.cn

## Abstract

Recent advancements in neural speech synthesis technologies have brought about widespread applications but have also raised concerns about potential misuse and abuse. Addressing these challenges is crucial, particularly in the realms of forensics and intellectual property protection. While previous research on source attribution of synthesized speech has its limitations, our study aims to fill these gaps by investigating the identification of sources in synthesized speech. We focus on analyzing speech synthesis model fingerprints in generated speech waveforms, emphasizing the roles of the acoustic model and vocoder. Our research, based on the multi-speaker LibriTTS dataset, reveals two key insights: (1) both vocoders and acoustic models leave distinct, model-specific fingerprints on generated waveforms, and (2) vocoder fingerprints, being more dominant, may obscure those from the acoustic model. These findings underscore the presence of model-specific fingerprints in both components, suggesting their potential significance in source identification applications.

## 1 Introduction

Recently, neural speech synthesis (also known as text-to-speech, or TTS) systems have made substantial advancements in generating highly realistic speech waveforms. These neural TTS systems can mainly be divided into two categories. The first category, pipeline systems, generally comprise an acoustic model such as Tacotron 2 (Shen et al., 2018), FastSpeech 2 (Ren et al., 2021) and GradTTS (Popov et al., 2021), alongside neural vocoders such as Parallel WaveGAN (Yamamoto et al., 2020), HiFiGAN (Kong et al., 2020), Style MelGAN (Mustafa et al., 2021) and Multiband MelGAN (Yang et al., 2020). The second category, end-to-end systems, also have been gaining traction in recent years. These systems, such as FastSpeech 2s (Ren et al., 2021) and VITS (Kim et al., 2021), are capable of generating speech waveforms directly from text.

While these advancements have found widespread application, they raise concerns about misuse in fraudulent activities, threatening security and privacy. Consequently, researchers have aimed to develop robust methods for differentiating genuine human speech from synthetic counterparts (Yi et al., 2023b). Initiatives like the ASVspoof challenges (Wu et al., 2015; Kinnunen et al., 2017; Wang et al., 2020; Yamagishi et al., 2021) and the ADD challenges (Yi et al., 2022) showcase advancements in audio anti-spoofing, including the use of front-end spectral features like linear frequency cepstral coefficients (LFCCs) (Todisco et al., 2018) as well as backend classification models like Res2Net (Gao et al., 2021) in anti-spoofing.

©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1160-1171, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers

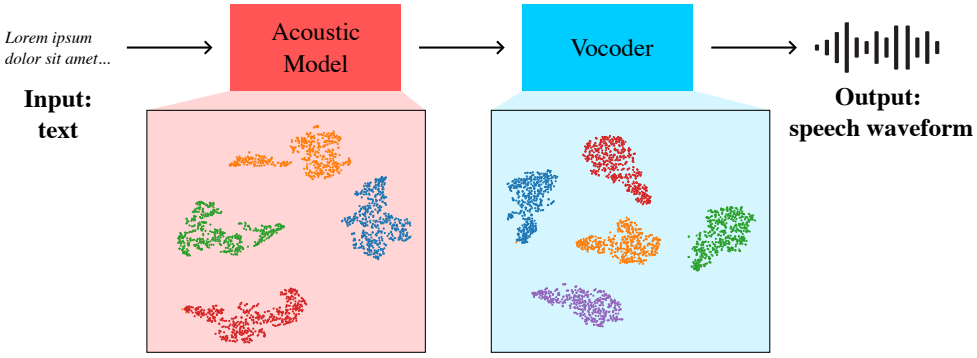



Figure 1: TTS pipeline and t-SNE projection of acoustic model and vocoder fingerprints.

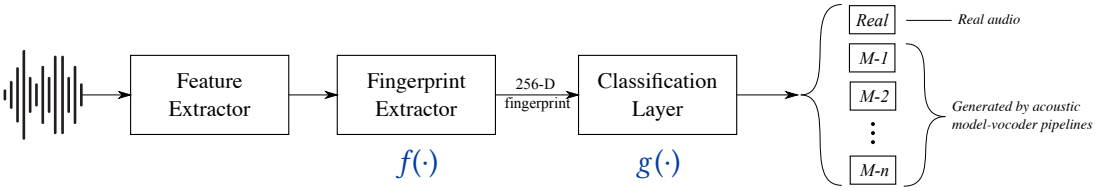



Figure 2: Fingerprint extraction pipeline.

There is a also growing interest in identifying the source of synthesized speech, especially in forensics and intellectual property protection. The ADD 2023 challenge (Yi et al., 2023a) includes a deepfake algorithm recognition track, addressing the recognition of synthesis tools, where various methods have been proposed (Lu et al., 2023; Tian et al., 2023). It is however worth noting that deepfake source attribution is a research topic that has not been extensively explored. Pons *et al.* (Pons et al., 2021) focused on upsampling artifacts but limited their investigation to visual representation without exploring exploitability for attribution. Similarly, Yan *et al.* (Yan et al., 2022) studied vocoder fingerprints but overlooked finer parameter differences within vocoders of the same architecture. This underscores the need for more comprehensive investigations covering various factors to enhance our understanding of deepfake source attribution. Notably, the traditional approach to TTS usually involves a pipeline system, where an acoustic model generates mel spectrograms from text, which are then converted into speech waveforms by a vocoder; mainstream approaches to source attribution have largely overlooked the potential of these components to leave fingerprints on the synthesized speech by performing attribution on a pipeline-level basis.

In the interest of filling this gap and fostering a more robust and nuanced comprehension of deepfake source attribution, we present our study. We aim to extract fingerprints from synthesized speech to identify their sources, focusing on acoustic model and vocoder fingerprints in pipeline TTS systems. In particular, we put forth and seek to explain the following questions: (1) Do acoustic models and vocoders leave model-specific fingerprints on the speech waveforms they generate? (2) When taking into consideration both the acoustic model and the vocoder, does one overshadow the other in terms of fingerprints?

To answer these questions, we train various acoustic models and vocoders on the LibriTTS (Zen et al., 2019) dataset to form TTS pipelines. Through waveform fingerprint extraction from these pipelines, we assess our hypothesis that both the acoustic model and vocoder leave model-specific fingerprints on the generated waveforms. Our experiments confirm these fingerprints' existence, highlighting the vocoder's more prominent role, potentially masking acoustic model fingerprints. These findings suggest the potential utility of model-specific fingerprints in source identification applications.

In this paper, we present our methodology in Section 2, followed by our experimental setup and results in Section 3, and then conclude with a discussion of our findings in Section 4.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1160-1171, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

## 2 Methodology

In this section we present the methodology employed for the verification of fingerprints within synthesized speech generated by TTS pipelines. Abstractly, a speech synthesis pipeline composed of a given acoustic model $a(\cdot)$ and a given vocoder $v(\cdot)$ can be viewed as a function that maps a text sequence $x$ to a speech waveform $y$, i.e., $y = v(a(x))$. Suppose there theoretically exist a "perfect" acoustic model $a^*$ and a "perfect" vocoder $v^*$ that generate "perfect" speech waveforms that are mathematically indistinguishable from genuine human speech. As current TTS pipelines are imperfect, with imperfections arising from both the theoretical foundation and the practical implementation, we can therefore expect that the speech waveforms generated by these pipelines will deviate from perfection; i.e.,

$$\begin{aligned} a(x) &= a^*(x) + \epsilon_a(x) \\ v(a(x)) &= v^*(a(x)) + \epsilon_v(a(x)) \end{aligned} \tag{1}$$

where $\epsilon_a$ and $\epsilon_v$ are residuals that contain the fingerprints of the acoustic model and the vocoder, respectively.

Inspired by the assumption in Marra et al. (2019), we assume that the fingerprints of the acoustic model and the vocoder are independent of each other, and that each of the residuals $\epsilon_a$ and $\epsilon_v$ contains the fingerprint plus a Gaussian noise, such that, with a sufficiently large dataset, the Gaussian noise can be ignored. We therefore seek to verify the existence of such fingerprints through an extraction network $f(\cdot)$ that extracts the fingerprints from the speech waveform $y$, i.e., either

$$\begin{aligned} f_a(y) &= \hat{\epsilon}_a(x) \\ f_v(y) &= \hat{\epsilon}_v(a(x)) \end{aligned} \tag{2}$$

as the case may be. We then train a classifier $g(\cdot)$ to classify the fingerprints extracted by $f(\cdot)$ into the corresponding acoustic model or vocoder, i.e.,

$$\begin{aligned} g_a(f(y)) &= a \\ g_v(f(y)) &= v \end{aligned} \tag{3}$$

The entire extraction and classification pipeline is shown in 2. We then investigate the performance of the classifier $g(\cdot)$ to determine whether the fingerprints of the acoustic model and the vocoder are distinguishable.

Following the methodology outlined above, we investigate the fingerprints of the acoustic model and the vocoder separately, and then investigate the relationship between the two fingerprints. We also investigate the robustness of the fingerprints against perturbations in the input speech samples.

### 2.1 Vocoder fingerprint

For vocoders, each model is characterized not only by the model architecture, but also by the precise weights in the neural network. This is especially true with generative adversarial networks (GANs) due to the adversarial nature of their training, meaning that the precise weights are sensitive to training setups (Brock et al., 2019). For this reason, we investigate the impact of both model architectures and training setups in vocoder fingerprints.

To study the impact of model architectures, we train vocoders of 4 architectures: Parallel WaveGAN (Yamamoto et al., 2020), HiFiGAN (Kong et al., 2020), Multiband MelGAN (Yang et al., 2020), and Style MelGAN (Mustafa et al., 2021), and investigate the fingerprints of the generated waveforms to determine whether the architecture leaves a fingerprint. To investigate the impact of training setup, we train several vocoders of the 4 aforementioned architectures, each with varying training setups (see Table 1), and then investigate the fingerprints of various vocoders trained with different setups to analyze the influence of training setup on the fingerprint.

CCL 2024

| Architecture | Model ID | Seed | Batch size |
|---|---|---|---|
| Parallel WaveGAN (PWG) | P0 | 1000 | 6 |
| | P1 | 1001 | 6 |
| | P2 | 1000 | 8 |
| | P3 | 1002 | 6 |
| | P4 | 1003 | 6 |
| HiFiGAN (HFG) | H0 | 1000 | 16 |
| | H1 | 1001 | 16 |
| | H2 | 1000 | 8 |
| Multiband MelGAN (MMG) | M0 | 1000 | 16 |
| | M1 | 1001 | 16 |
| | M2 | 1000 | 8 |
| StyleMelGAN (SMG) | S0 | 1000 | 16 |
| | S1 | 1001 | 16 |
| | S2 | 1000 | 8 |

Table 1: Vocoder model setups.

| Exp. | Training set | Validation set | Test set |
|---|---|---|---|
| V1 | P0, H0, M0, S0 | (same methods as training set) | (same methods as training set) |
| V2 | P0, P1, P3, P4 | (same methods as training set) | (same methods as training set) |
| V3 | P0, H0, M0, S0 | P1, P2, H1, H2, M1, M2, S1, S2 | P1, P2, H1, H2, M1, M2, S1, S2 |
| A1 | F2+H0, GD+H0, T2+H0 | (same methods as training set) | (same methods as training set) |
| A2 | F2+H0, F2+P0, F2+S0, GD+H0, GD+P0, GD+S0, T2+H0, T2+P0, T2+S0 | (same methods as training set) | (same methods as training set) |
| R1 | F2+H0, GD+H0, T2+H0 | F2+P1, F2+S1, GD+P1, GD+S1, T2+P1, T2+S1 | F2+P1, F2+S1, GD+P1, GD+S1, T2+P1, T2+S1 |
| R2 | T2+P0, T2+H0, T2+M0, T2+S0 | F2+P0, F2+H0, F2+M0, F2+S0 | F2+P0, F2+H0, F2+M0, F2+S0 |
| N1 | (R2 training set) | (R2 validation set) | (R2 test set with noise at 10 dB SNR) |
| N2 | (R2 training set) | (R2 validation set) | (R2 test set with reverberation) |
| N3 | (R2 training set) | (R2 validation set) | (R2 test set with speed adjustment) |

Table 2: Fingerprint analysis experiment setups. (Exp. = Experiment ID)
Acoustic models used: FastSpeech 2 (F2), Grad-TTS (GD), Tacotron 2 (T2)
Vocoders used: see Table 1
In Experiments V1, V2, A1, and A2, different inputs are used with the same set of vocoders/acoustic models to generate samples for the training, validation and test sets.

### 2.2 Acoustic model fingerprint

Acoustic models are paired with a vocoder to generate waveforms, therefore, we study two aspects of the acoustic model fingerprint: when one vocoder is used to generate waveforms from the mel spectrograms produced by the acoustic model, and when multiple vocoders are used to generate waveforms from the mel spectrograms produced by the acoustic model. We train acoustic models of 3 architectures: Tacotron 2 (Shen et al., 2018), FastSpeech 2 (Ren et al., 2021), and GradTTS (Popov et al., 2021), and use our trained models in conjunction with the previously trained vocoders to generate waveforms from the mel spectrograms produced by the acoustic models. We then investigate the fingerprints of the generated waveforms to determine whether the acoustic model leaves a fingerprint.

### 2.3 Relationship between fingerprints

When it comes to the relationship between vocoder and acoustic model fingerprints, there are two competing hypotheses, both needing to be verified: either (1) vocoder fingerprints can visibly interfere with acoustic model fingerprints, or (2) acoustic model fingerprints can visibly

CCL 2024

interfere with vocoder fingerprints. To verify both hypotheses, we design two experiments with complementary setups, aimed to verify the two aforementioned, somewhat contradictory hypotheses. By setting one module — acoustic model or vocoder, as the case may be — to be known in the testing set, and testing the fingerprint against unknown vocoders or acoustic models, respectively, we can verify the two hypotheses. The setups of the two experiments are shown in Table 2, in rows R1 and R2.

### 2.4 Perturbations

We furthermore seek to investigate the robustness of these fingerprints against perturbations in speech samples, namely noise, reverberation, and speed adjustment. To this end, we perform three experiments, for each of the aforementioned perturbations, where we add the specific perturbation to the speech samples in the test set, and then investigate the fingerprints of the generated waveforms to determine the extent of impact such perturbation would have on the original fingerprints. The setups of the three experiments are shown in Table 2, in rows N1, N2 and N3.

## 3 Experiments and results

In this section, we outline our conducted experiments and the subsequent findings. We have undertaken a series of ten distinct experiments. The first five experiments are conducted to verify the existence of vocoder (V1–V3) and acoustic model (A1–A2) fingerprints in synthesized speech, and the next two (R1–R2) are conducted to verify the influence of the acoustic model and the vocoder on the fingerprints. Finally, three experiments (N1–N3) are conducted to investigate the robustness of the fingerprints against perturbations in the input speech samples. The setups of the experiments are shown in Table 2.

### 3.1 Experimental setup

In the process of analyzing speech waveforms, each sample undergoes a detailed extraction method to obtain its linear frequency cepstral coefficients (LFCCs). This process begins with the segmentation of the speech waveform into frames, from which we extract 20 LFCCs for each frame. This extraction uses a Hamming window, which is a specific type of function used to smooth out the signal, minimizing the edge effects during the analysis. The parameters for this window are carefully chosen, with a window length set at $M = 480$ and hop length of $H = 240$.

We standardize the size of the resulting LFCC matrix across all samples, since the matrix might vary in the number of frames depending on the duration of the speech sample. To address this variance, we either truncate the matrix to contain exactly 500 frames if it exceeds this number, or we pad it with zeros to reach the frame count of 500 if it contains fewer frames. This step is crucial for maintaining consistency across the dataset, allowing the neural network to process the data more effectively.

Inspired by Marra et al. (2019), the LFCC representation of the speech sample is then passed to a Res2Net network (Gao et al., 2021) and projected onto a 256-dimension feature space, then passed through a single fully-connected (FC) layer classifier to decide between the finite number of sources. The entire network, i.e., both the extraction network and the classifier, is trained using cross-entropy loss:

$$\mathcal{L} = -\sum_{i=1}^{N}\sum_{j=1}^{C} y_i^{(j)} \log(p_i^{(j)}) \tag{4}$$

where $N$ is the number of samples, $C$ is the number of classes, $y$ and $p$ denote the ground truth label and the predicted probability, respectively.

### 3.2 Dataset

In our experiments, we use the multi-speaker dataset LibriTTS (Zen et al., 2019) to both train acoustic models and vocoders as well as to generate the waveform corpus for fingerprint

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1160-1171, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

| Variable | Exp. | Precision | Recall | $F_1$ |
|---|---|---|---|---|
| Vocoder | V1 | 98.46 | 98.44 | 98.44 |
| | V2 | 95.74 | 95.73 | 95.73 |
| | V3 | 91.57 | 91.19 | 91.20 |
| Acoustic model | A1 | 99.79 | 99.79 | 99.79 |
| | A2 | 98.99 | 98.99 | 98.99 |
| Relationship | R1 | 23.75 | 25.01 | 10.05 |
| | R2 | 98.46 | 98.42 | 98.42 |
| Perturbation | N1 | 98.21 | 98.16 | 98.17 |
| | N2 | 87.82 | 86.36 | 86.03 |
| | N3 | 91.24 | 91.12 | 91.07 |

Table 3: Summary of experiment results (%)

extraction. The "clean" training set of LibriTTS, containing 160,369 genuine utterances from 1,230 speakers, totaling 262 hours of speech, is used to train all models until convergence.

In departure from the traditional setup where the validation set and training set are from the same distribution, we synthesize the validation set using the same setups as the test set (see Table 2), since we are interested in the performance of the classifier on the test set, which is synthesized using different setups from the training set. The training, validation and test sets are, however, disjoint, to help avoid overfitting.

### 3.3 Evaluation metric

The performance of fingerprint extraction and analysis should align with the ability to classify the audio samples and identify their source of generation. To thoroughly evaluate the performance of our fingerprint extraction model, we employ a trio of established metrics commonly utilized in classification tasks: precision, recall, and the F1 score. These metrics offer a comprehensive view of the model's accuracy and its ability to distinguish between different sources of generated speech. Beyond these numerical metrics, we delve into the analysis of the confusion matrix associated with our classifier, to gain insight into the performance of the fingerprint extraction model, as it visualizes the performance of a classifier by identifying the number of instances where the classifier misclassified the speech samples and in turn evaluate the performance of the fingerprint extraction model, on a per-category basis. This allows us to go beyond the global metrics and gain a more nuanced understanding of the model's performance.

### 3.4 Results and analysis

Following the methodology outlined in Section 2, and the setup of the experiments in Table 2, we conduct a series of experiments, aimed at verifying the existence of vocoder and acoustic model fingerprints in synthesized speech, and to investigate the relationship between the two fingerprints. We also investigate the robustness of the fingerprints against perturbations in the input speech samples. The results of these experiments are summarized in Table 3, and the confusion matrices of the experiments are shown in Figure 3.

#### 3.4.1 Vocoder fingerprints

For vocoder fingerprint analysis, we use copy synthesis to emulate a fixed acoustic model, minimizing variables. We train 3 vocoders for each of the four architectures: Parallel WaveGAN (PWG), HiFiGAN (HFG), Multiband MelGAN (MMG), and Style MelGAN (SMG), each with varying seeds and batch sizes (see Table 1). We additionally train two more PWG vocoders with different initialization seeds to investigate the impact of training setup on fingerprints. Three experiments assess vocoder architecture and training setup impact on fingerprints. In V1 and V3, we apply copy synthesis on the training set using P0, H0, and M0 models, labeling results by vocoder architecture for Res2Net classifier training. The validation and testing sets differ in vocoder models used for copy synthesis (see Table 2). In V2 we investigate training setup

CCL 2024

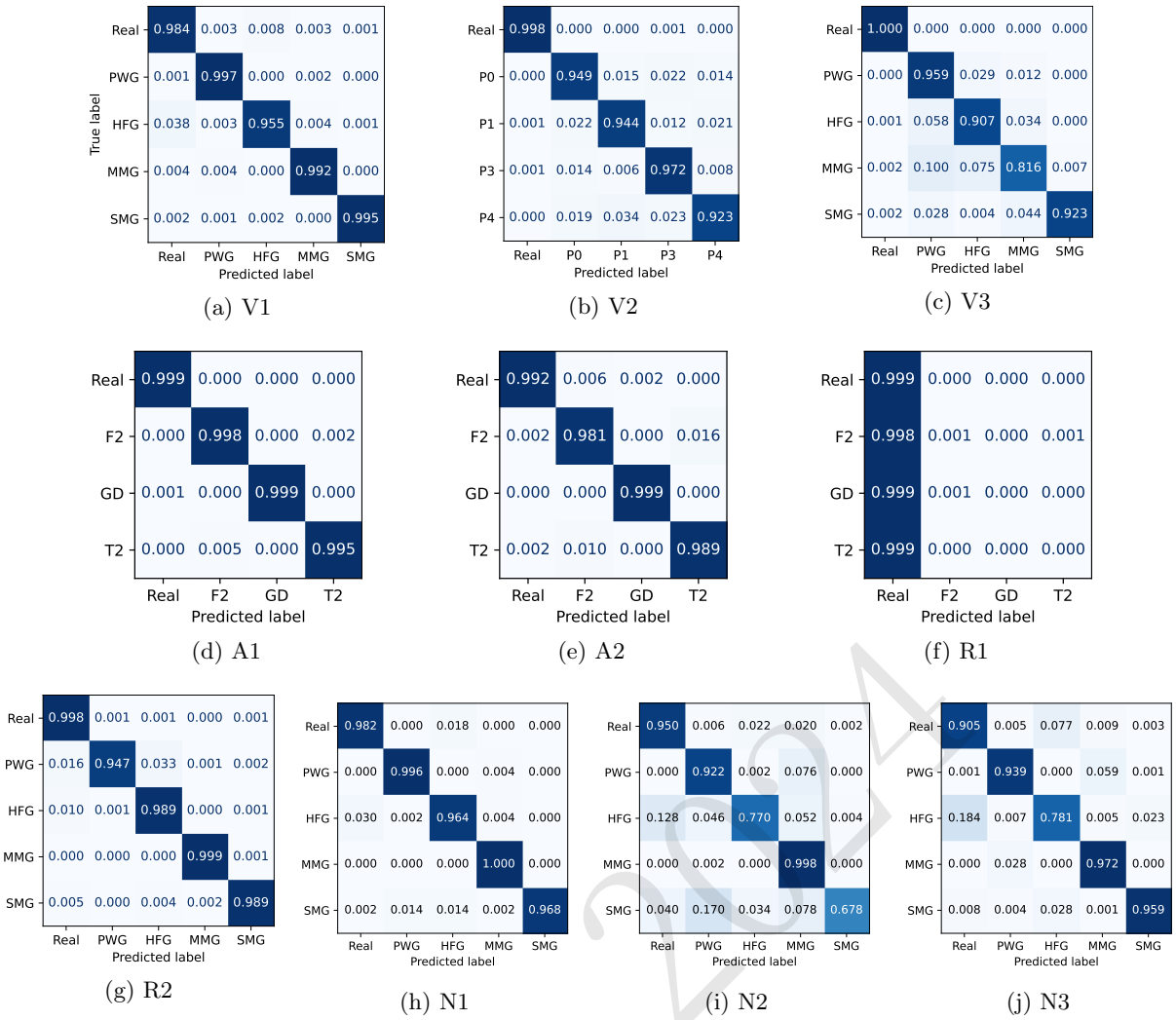



Figure 3: Confusion matrices of experiments in our study.

impact by classifying along initialization seed lines within the same architecture (i.e., Parallel WaveGAN).

Results are summarized in the V1–V3 rows of Table 3 as well as the confusion matrices in Figures 3a to 3c. The high F1 score of V1 and the highly diagonal confusion matrix in Figure 3a confirm architecture-specific fingerprints in the generated waveforms serving as strong indicators of source. In contrast, the slight degredation of classification performance in Experiment V3, as shown in Figure 3c, suggests vocoder training setups weaken fingerprints, an observation corroborated by Experiment V2 (Figure 3b). Overall, results confirm the existence of finer-grained vocoder fingerprints based on parameter differences, but with less pronounced impact than architectural ones.

### 3.4.2 Acoustic model fingerprints

We investigate and analyze the fingerprint of the acoustic model. Three models, Tacotron 2, FastSpeech 2 and Grad-TTS, are trained on the LibriTTS dataset. These models generate labeled speech waveforms from the LibriTTS training set text corpus. We use the entire speech corpus, encompassing both real and synthesized samples, to train the Res2Net classifier. The same approach applies to validation and test sets. To thoroughly examine the detectability and characteristics of the acoustic model fingerprint, we conduct two distinct experiments. Experiment A1 serves as a baseline, where we limit the diversity of the vocoders used in the synthesis process. Experiment A2, on the other hand, introduces a wider variety of vocoders into the mix

CCL 2024

(see Table 2). This variation allows us to assess the impact of vocoder diversity on the ability of the Res2Net classifier to learn and identify acoustic model fingerprints, and more importantly, to ascertain that the fingerprint features learned by our model are indeed acoustic model-specific.

The results of our study are encapsulated in confusion matrices, as referenced in our report (Figures 3d to 3e), and the detailed analysis is provided in rows A1–A2 of Table 3. These results shed light on our question, demonstrating that the acoustic model fingerprint is indeed detectable with a high degree of accuracy when the vocoder used during the synthesis process is either fixed or known. This revelation underscores the effectiveness of a fixed vocoder setup in enhancing the learning of acoustic model fingerprints. However, it is also noteworthy that introducing a variance in the vocoders, as done in Experiment A2, does not significantly impede the learning process. This is particularly true when the vocoder variance is evenly distributed across the categories under investigation. The insights gained from Experiment A2 highlight the resilience of the acoustic model fingerprint learning process, even in the face of increased diversity in vocoder usage.

### 3.4.3 Relationship between fingerprints

As outlined in Section 2.3, we investigate the relationship between vocoder and acoustic model fingerprints with a pair of complementary experiments, R1 and R2. In R1, we fix the acoustic model and vary the vocoder; the acoustic models used to generate the testing set are known, but the vocoder (H0) is swapped for two unknown models (P1 and S1). In R2, we fix the vocoder architectures, but vary the acoustic models used to generate the testing set (TN being replaced by FS). The results of these experiments are summarized in the R1–R2 rows of Table 3 and the confusion matrices in Figures 3f to 3g. Results in Figure 3f demonstrate that when vocoders vary, the acoustic model fingerprint becomes nearly undetectable, aligning with our hypothesis that vocoder fingerprints overshadow acoustic model fingerprints. This is further supported by the results in Figure 3g, where the vocoder fingerprint is still detectable when the acoustic model varies.

The experiment outcomes align with our expectations, considering the distinct roles of the vocoder and acoustic model in the process of speech synthesis. The vocoder directly generates the waveform by translating the spectrogram produced by the acoustic model into speech. Given its role as the last step in the synthesis chain, the vocoder's involvement in the generation of the speech waveform is direct and substantial. It takes the more compact, frequency-based information from the spectrogram and transforms it into the waveform that we ultimately hear as speech. This direct engagement implies that it has a significant influence on the final characteristics of the speech waveform. Consequently, it's plausible to suggest that the vocoder leaves a more distinguishable fingerprint on the generated speech than the acoustic model does.

This distinction between the vocoder and acoustic model's contributions is essential for understanding how different components of the speech synthesis process affect the final product. It implies that when analyzing synthesized speech for the purpose of identifying its source or assessing its quality, the vocoder's characteristics may be a more telling factor than those of the acoustic model. This insight is crucial for the development of speech synthesis systems and the evaluation of their performance.

## 3.5 Fingerprint robustness against speech perturbation

Given the analyses above, demonstrating the existence of distinct fingerprints for both vocoders and acoustic models, as well as the overshadowing of acoustic model fingerprints by vocoder fingerprints, we further investigate the robustness of the fingerprints against perturbations in the input speech samples. We analyze the robustness from three aspects:

1. Noise: We add white noise to the speech samples in the test set of R2, with various signal-to-noise ratios (SNR) of 0–10 dB.
2. Reverberation: We add reverberation to the speech samples in the test set of R2, with various reverberation times (RT) of 0.5–1.5 seconds.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1160-1171, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

| Variable | Param. | Precision | Recall | $F_1$ |
|---|---|---|---|---|
| Noise | SNR=10dB | 98.21 | 98.16 | 98.17 |
| | SNR=5dB | 97.04 | 96.98 | 97.00 |
| Reverberation | RT=0.5s | 87.82 | 86.36 | 86.03 |
| | RT=1.0s | 85.12 | 83.68 | 83.35 |
| Speed adjustment | 0.9x | 91.24 | 91.12 | 91.07 |
| | 1.1x | 90.12 | 89.98 | 89.94 |

Table 4: Summary of perturbation experiment results (%)

3. Speed adjustment: We adjust the speed of the speech samples in the test set of R2, with various speed factors of 0.9x–1.1x.

We conduct an experiment where we add noise to the speech samples in the test set. We use SoX[0] to add white noise to the speech samples in the test set of R2, with a signal-to-noise ratio (SNR) of 10 dB. We then investigate the fingerprints of the generated waveforms to determine the extent of impact such perturbation would have on the original fingerprints, as summarized in Row N1 of Table 3.

Similar setups are used to investigate the robustness of the fingerprint extraction model against reverberations and speed adjustments (varying from 0.9x to 1.1x), with results listed in rows N2 and N3 of Table 3, respectively. The results of these experiments show that the fingerprint extraction model is somewhat robust against perturbations in the input speech samples, with a slight degradation in performance. Among the three perturbations tested, reverberations seem to have the most adverse effect on the fingerprints. This is likely due to the fact that reverberations are more likely to affect the spectral characteristics of the speech samples, which are the basis of the fingerprints, than noise and speed adjustments.

We furthermore conduct experiments with various signal-to-noise ratios (SNR) and reverberation times (RT) to investigate the robustness of the fingerprints against noise and reverberation. The results of these experiments are summarized in Table 4. The results reinforce our analysis of the fingerprint robustness against perturbations, with speed adjustments and reverberation having a more pronounced impact on the fingerprints than noise, with reverberation being the most detrimental to the fingerprints.

### 3.6 Visualization

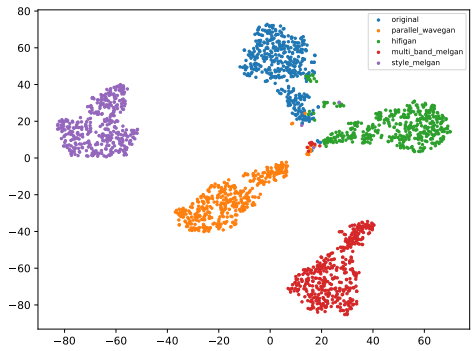
Given the results reported in the previous section, to better and more intuitively understand the fingerprints of the vocoders and acoustic models, we visualize the fingerprints of the vocoders and acoustic models using t-SNE (van der Maaten and Hinton, 2008).

We focus first on the vocoders. By applying t-SNE, we project the high-dimensional fingerprints of various vocoders onto a two-dimensional plane. This visualization is reflected in Figure 4a. The resulting projection clearly illustrates how the fingerprints of the vocoders are dispersed, forming clearly distinguishable clusters on the plane. Each of these clusters represents a unique vocoder architecture, unequivocally demonstrating the presence of distinct, identifiable fingerprints for each vocoder.

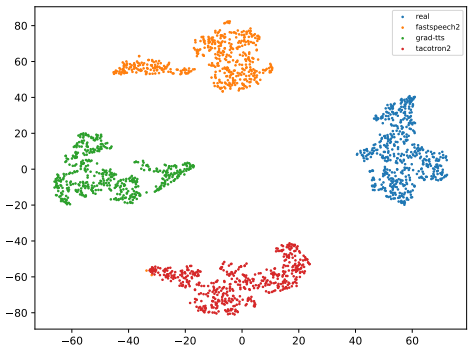
Following the analysis of vocoders, we extend the same t-SNE visualization technique to the acoustic models. The outcomes of this process are captured in Figure 4b, where, akin to the vocoders, the fingerprints of the acoustic models are also neatly organized into discernible clusters. This pattern reinforces the premise that acoustic models, much like vocoders, possess their own unique fingerprints. The clarity with which these fingerprints can be distinguished affirms the potential for precise identification and differentiation of acoustic models based on their inherent characteristics.

Figure 4c shows the t-SNE projection of the fingerprints of the vocoders and acoustic models

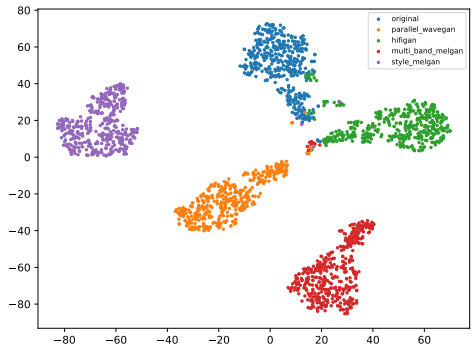
[0]Available at https://sourceforge.net/projects/sox/



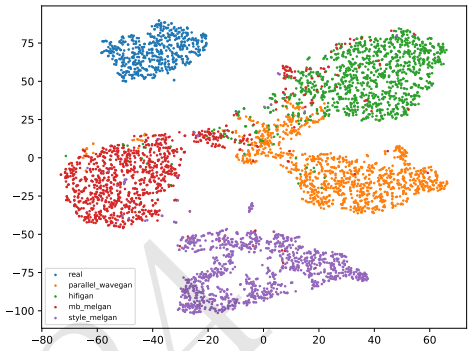

(a) Vocoder fingerprints (V3)




(b) Acoustic model fingerprints (A2)




(c) Relationship between vocoder and acoustic model fingerprints (R2)




(d) Perturbation experiment: audio speed readjustment (N3)

Figure 4: t-SNE representation of fingerprints extracted in our experiments.

in Experiment R2. As with the previous visualizations, the fingerprints of the vocoders are clearly separated into distinct clusters, despite the presence of different and unknown acoustic models. This result confirms that the vocoder fingerprints are more prominent than the acoustic model fingerprints, potentially overshadowing the acoustic model fingerprints.

Lastly, Figure 4d shows the t-SNE projection of the fingerprints of the perturbation experiment N3. The fingerprints of the perturbed speech samples are still distinguishable, albeit with a slight overlap, indicating that the fingerprint extraction model is somewhat robust against perturbations in the input speech samples.

## 4 Conclusion

In conclusion, our study centers on extracting fingerprints from acoustic models and vocoders to identify the sources of synthesized speech. Our experiments confirm the existence of distinct, model-specific fingerprints for both vocoders and acoustic models. Notably, we observe that vocoders tend to leave more prominent fingerprints, potentially overshadowing those of the acoustic models. This discovery holds promise, indicating the feasibility of employing fingerprint extraction for source attribution in synthesized speech. This insight is particularly relevant for researchers and developers in the field of computational linguistics and speech synthesis, as it highlights the importance of carefully selecting or designing vocoders to achieve the desired qualities in synthesized speech. Nonetheless, our current research leaves room for further exploration. We have yet to consider the potential impact of different languages, which entail diverse phonological feature distributions, on the fingerprints of generated waveforms. It will also be interesting and important to devise methods for extracting fingerprints from end-to-end TTS systems, which are gaining traction in recent years. Lastly, even in the context of pipeline TTS systems, we have yet to come up with a reliable method for extracting acoustic-model-specific fingerprints despite the masking presence of vocoder fingerprints. These unexplored dimensions

CCL 2024

will form the focal point of our future investigations, seeking to provide a more comprehensive understanding of the nuances in the fingerprints of synthesized speech.

## Acknowledgements

This work is supported by the National Natural Science Foundation of China (NSFC) (No. 62322120, No. 62306316, No U21B2010, No. 62206278), as well as the ANSO Scholarship for Young Talents.

## References

Andrew Brock, Jeff Donahue, and Karen Simonyan. 2019. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*.

Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. 2021. Res2Net: A new multi-scale backbone architecture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(2):652–662.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *Proceedings of the 38th International Conference on Machine Learning*, page 5530–5540. PMLR.

Tomi Kinnunen, Md. Sahidullah, Héctor Delgado, Massimiliano Todisco, Nicholas Evans, Junichi Yamagishi, and Kong Aik Lee. 2017. The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection. In *Interspeech 2017*, page 2–6. ISCA.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. In *Advances in Neural Information Processing Systems*, volume 33, pages 17022–17033.

Jingze Lu, Yuxiang Zhang, Zhuo Li, Zengqiang Shang, WenChao Wang, and Pengyuan Zhang. 2023. Detecting unknown speech spoofing algorithms with nearest neighbors. In *Proceedings of IJCAI 2023 Workshop on Deepfake Audio Detection and Analysis*.

Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. 2019. Do GANs Leave Artificial Fingerprints? In *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 506–511. IEEE Computer Society.

Ahmed Mustafa, Nicola Pia, and Guillaume Fuchs. 2021. Stylemelgan: An efficient high-fidelity adversarial vocoder with temporal adaptive normalization. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6034–6038.

Jordi Pons, Santiago Pascual, Giulio Cengarle, and Joan Serrà. 2021. Upsampling artifacts in neural audio synthesis. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3005–3009.

Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. 2021. Grad-TTS: A diffusion probabilistic model for text-to-speech. In *Proceedings of the 38th International Conference on Machine Learning*, page 8599–8608. PMLR.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2021. FastSpeech 2: Fast and high-quality end-to-end text to speech. In *International Conference on Learning Representations (ICLR)*.

Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, Rif A. Saurous, Yannis Agiomvrgiannakis, and Yonghui Wu. 2018. Natural TTS synthesis by conditioning Wavenet on Mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783.

Ye Tian, Yunkun Chen, Yuezhong Tang, and Boyang Fu. 2023. Deepfake algorithm recognition through multi-model fusion based on manifold measure. In *Proceedings of IJCAI 2023 Workshop on Deepfake Audio Detection and Analysis*.

Massimiliano Todisco, Héctor Delgado, Kong Aik Lee, Md Sahidullah, Nicholas Evans, Tomi Kinnunen, and Junichi Yamagishi. 2018. Integrated presentation attack detection and automatic speaker verification: Common features and gaussian back-end fusion. In *Proc. Interspeech 2018*, pages 77–81.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605.

Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Héctor Delgado, Andreas Nautsch, Nicholas Evans, Md Sahidullah, Ville Vestman, Tomi Kinnunen, Kong Aik Lee, Lauri Juvela, Paavo Alku, Yu-Huai Peng, Hsin-Te Hwang, Yu Tsao, Hsin-Min Wang, Sébastien Le Maguer, Markus Becker, Fergus Henderson, Rob Clark, Yu Zhang, Quan Wang, Ye Jia, Kai Onuma, Koji Mushika, Takashi Kaneda, Yuan Jiang, Li-Juan Liu, Yi-Chiao Wu, Wen-Chin Huang, Tomoki Toda, Kou Tanaka, Hirokazu Kameoka, Ingmar Steiner, Driss Matrouf, Jean-François Bonastre, Avashna Govender, Srikanth Ronanki, Jing-Xuan Zhang, and Zhen-Hua Ling. 2020. ASVspoof 2019: a large-scale public database of synthesized, converted and replayed speech. *Computer Speech & Language*, 64:101–114.

Zhizheng Wu, Tomi Kinnunen, Nicholas Evans, Junichi Yamagishi, Cemal Hanilçi, Md. Sahidullah, and Aleksandr Sizov. 2015. ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge. In *Interspeech 2015*, page 2037–2041. ISCA.

Junichi Yamagishi, Xin Wang, Massimiliano Todisco, Md Sahidullah, Jose Patino, Andreas Nautsch, Xuechen Liu, Kong Aik Lee, Tomi Kinnunen, Nicholas Evans, et al. 2021. Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection. *arXiv preprint arXiv:2109.00537*.

Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. 2020. Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6199–6203.

Xinrui Yan, Jiangyan Yi, Jianhua Tao, Chenglong Wang, Haoxin Ma, Tao Wang, Shiming Wang, and Ruibo Fu. 2022. An initial investigation for detecting vocoder fingerprints of fake audio. In *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia (DDAM '22)*, page 61–68.

Geng Yang, Shan Yang, Kai Liu, Peng Fang, Wei Chen, and Lei Xie. 2020. Multi-band MelGAN: Faster waveform generation for high-quality text-to-speech. *arXiv:2005.05106 [cs, eess]*.

Jiangyan Yi, Ruibo Fu, Jianhua Tao, Shuai Nie, Haoxin Ma, Chenglong Wang, Tao Wang, Zhengkun Tian, Ye Bai, Cunhang Fan, et al. 2022. ADD 2022: the first audio deep synthesis detection challenge. In *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 9216–9220. IEEE.

Jiangyan Yi, Jianhua Tao, Ruibo Fu, Xinrui Yan, Chenglong Wang, Tao Wang, Chu Yuan Zhang, Xiaohui Zhang, Yan Zhao, Yong Ren, Le Xu, Junzuo Zhou, Hao Gu, Zhengqi Wen, Shan Liang, Zheng Lian, Shuai Nie, and Haizhou Li. 2023a. ADD 2023: the second audio deepfake detection challenge. In *Proceedings of IJCAI 2023 Workshop on Deepfake Audio Detection and Analysis*.

Jiangyan Yi, Chenglong Wang, Jianhua Tao, Xiaohui Zhang, Chu Yuan Zhang, and Yan Zhao. 2023b. Audio deepfake detection: A survey. *arXiv 2308.14970*.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. LibriTTS: A corpus derived from librispeech for text-to-speech. In *Proc. Interspeech 2019*, pages 1526–1530.

CCL 2024

# Knowledge Graph-Enhanced Recommendation with Box Embeddings

**Qiuyu Liang**[1], **Weihua Wang**[1,2,3]*, **Lei Lv**[1], **Feilong Bao**[1,2,3]
[1] College of Computer Science, Inner Mongolia University, Hohhot, China
[2] National and Local Joint Engineering Research Center of Intelligent Information Processing Technology for Mongolian, Hohhot, China
[3] Inner Mongolia Key Laboratory of Mongolian Information Processing Technology, Hohhot, China
wangwh@imu.edu.cn

## Abstract

Knowledge graphs are used to alleviate the problems of data sparsity and cold starts in recommendation systems. However, most existing approaches ignore the hierarchical structure of the knowledge graph. In this paper, we propose a box embedding method for knowledge graph-enhanced recommendation system. Specifically, the box embedding represents not only the interaction between the user and the item, but also the head entity, the tail entity and the relation between them in the knowledge graph. Then the interaction between the item and the corresponding entity is calculated by the multi-task attention unit. Experimental results show that our method provides a large improvement over previous models in terms of Area Under Curve (AUC) and accuracy in publicly available recommendation datasets with three different domains.

## 1 Introduction

Recommendation systems have been proposed to solve the information explosion problem and to improve the user experience in various applications. Early recommendation systems were mainly based on collaborative filtering algorithms. These algorithms build user-item interaction matrices and perform similarity calculations to obtain final recommendations. However, collaborative filtering-based approaches encounter data sparsity and cold-start problems in practical applications.

To address these problems, some researchers have proposed recommendation methods that use Knowledge Graphs (KGs) as auxiliary information to improve the performance of recommendation tasks. Since KGs contain rich semantic relationships between entities corresponding to user-related items, the recommended items can take advantage of interactions with entities in the KG. Previous knowledge graph enhanced recommendation methods mainly used multiple modules to learn the representation of knowledge graph and recommendation task independently. For example, CKE (Zhang et al., 2016) used the TransR to encode the knowledge graph, while extracted the text and image features of the recommendation task through another self-encoding structure. Recently, several researchers have proposed methods for learning entities and items using multi-task learning. The core idea of such methods trained feature interaction layers jointly to connect the recommendation task and the knowledge graph-related task. For example, MKR (Wang et al., 2019) used cross units for training the knowledge graph embedding task and recommendation task. But the units were computed by matrix multiplication. This lead to the insufficient learning about the feature. In addition, few existing methods can model the semantic hierarchical structure relationship between recommendation data and the knowledge graph.

To address above shortcomings, we proposed **BoxMAKR**, a box embedding for knowledge graph-enhanced recommendation. Specifically, our model utilizes the attention mechanism to learn higher-order interaction features between items and entities more efficiently, and then embeds semantic hierarchical relationships into entities, users and items. We conduct experiments on three recommendation datasets, and the results show that BoxMAKR proposed in this paper achieves a large improvement compared with advanced models such as MKR.

* Corresponding author.
©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1172-1182, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

## 2 Related Work

**Knowledge graph enhanced recommendation systems.** The recommendation systems can provide users with interesting content based on their historical information. Early recommendation algorithms include collaborative filtering (Hu et al., 2008; Wang et al., 2020; Liu and Wang, 2022; Jiang et al., 2022), factorization machines (Rendle, 2010), and matrix factorization (Zhang et al., 2006). These algorithms basically used the historical interaction data to calculate the similarity between each user and item. Recently, some methods proposed more efficient mechanisms to learn the representation of users and items. However, these methods still suffered from data sparsity and cold start problems. For example, DNNRec (Kiran et al., 2020) integrated side information about users and items into a deep neural network to alleviate the cold start problem. DHMR (Khan et al., 2021) proposed higher-order non-linear latent feature to learn interactions from reviews and metadata information. But learning the lower-order feature interactions of users and items were separately.

Besides user reviews information, knowledge graphs also can be considered as an additional information source. This kind of task is called knowledge graph enhanced recommendation system. They are usually contain two subtasks: knowledge graph embedding and the recommendation task. To integrate the two tasks more effectively, more and more researchers focused on multi-task learning. For instance, MKR (Wang et al., 2019) designed an interaction unit to share the features between items and entities and trained jointly. Ripp-MKR (Wang et al., 2021) used the core idea of RippNet (Wang et al., 2018) to combine the knowledge graph with historical user interaction to represent user features based on MKR. TransMKR (Hu et al., 2022) improved the knowledge graph embedding module of MKR with TransR, which enhanced the representation capability of knowledge graph embedding.

**Geometric embedding**. The semantic hierarchical structure relationship of data explicitly modeled in the embedding space is very useful for solving natural language processing tasks. However, modelling the structure of word in the Euclidean space requires large embedding dimension and additional constraints. These constrain increased the computational complexity of the model. Therefore, some researchers proposed non-Euclidean embedding methods. For example, POE (Lai and Hockenmaier, 2017) was proposed to learn the sequential mapping of objects based on the semantic hierarchical structure instead of the traditional distance relationship. Its core idea is to model the objects using the partial order in the semantic hierarchical structure. Reference (Nickel and Kiela, 2017; Maximillian and Kiela, 2018; Liang et al., 2024) modeled objects' semantic hierarchical structure by similarity between the concepts in hyperbolic space. Box embeddings (Onoe et al., 2021; Patel and Sankar, 2020; Vilnis et al., 2018) used hyper-rectangles(box) to represent semantic hierarchical structure between objects as explicit conditional probabilities.

## 3 Problem Formula

In the recommendation task, there exists a user set $U = \{u_1, u_2, ..., u_m\}$, an item set $V = \{v_1, v_2, ..., v_n\}$, and a user-item interaction matrix $\mathcal{Y} \in \mathbb{R}^{m \times n}$, where $y_{uv} \in \{0, 1\}$. This user-item matrix represents whether an interaction exists between user $u_m$ and item $v_n$, which is defined by users' feedbacks. In the matrix $\mathcal{Y}$, $y_{uv} = 1$ represents the user $u$ engaged with item $v$, otherwise $y_{uv} = 0$. These feedbacks include implicit users' click, watching, browsing. Furthermore, a knowledge graph $\mathcal{G}$ can be composed of many triples $(h, r, t)$. Where $h$ denotes the head entity, $t$ denotes the tail entity, and $r$ denotes the relationship between them. In the knowledge graph each triple is associated with the items$v$. Also, the item $v \in V$ may be related to one or more entities in the knowledge graph.

Thus, the knowledge graph enhanced recommendation task can be formulated as follows: given a user set $U$, an item set $V$, a user-item interaction matrix $\mathcal{Y}$ and a knowledge graph $\mathcal{G}$, the model will predict the $\hat{y}_{uv}$ which means the user $u$ is interested in item $v$.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1172-1182, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

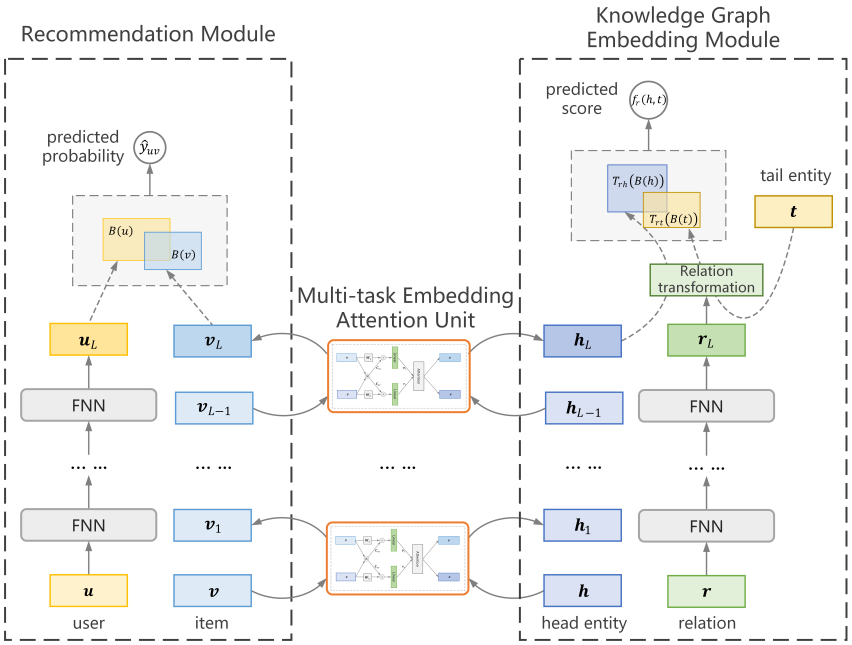



Figure 1: The architecture of BoxMAKR model.

# 4 Method

## 4.1 Model Overview

BoxMAKR consists of three parts: multi-task embedding attention units, recommendation module, and knowledge graph embedding module. The architecture of BoxMAKR is shown in Figure 1.

- The multi-task embedding attention units are implemented based on the attention mechanism for modeling the higher-order interaction features between items in the recommendation module and entities in the knowledge graph embedding module.
- The inputs of the recommendation module are users and items. Firstly, user and item embeddings are learned by multi-layer feedforward neural networks and multi-task embedding attention units. Then the user and item are learned by box embeddings. Finally, the prediction probabilities of users for items are output using condition probability of box embeddings.
- The knowledge graph embedding module has a similar network structure as the recommendation module. The relation and entity embeddings learned by multi-layer feedforward neural networks and multi-task embedding attention units. Then the head and tail entities are learned by box embeddings. Finally, the prediction scores of triples are output using condition probability of box embeddings.

## 4.2 Box Embeddings for Semantic Hierarchical Structure

Box embeddings represent each object as a hyper-rectangle $B(x)$ in the $d$ dimensional embedding space, each hyper-rectangle is called a box. The box is represented in each dimension by a pair of vectors $\langle \boldsymbol{x}_m, \boldsymbol{x}_M \rangle$. In each dimension $i \in \{1, 2, ..., d\}$ there is $x_{m,i} < x_{M,i}$, where $x_{m,i}$ and $x_{M,i}$ represent the vector values of the bottom-left and upper-right positions of the box, respectively, i.e., the minimum and maximum values of the box in the embedding space. The box embeddings are shown in Figure 2. If the box embeddings space $\Omega_B \subseteq \mathbb{R}^d$ is normalized to volume 1, then the box can

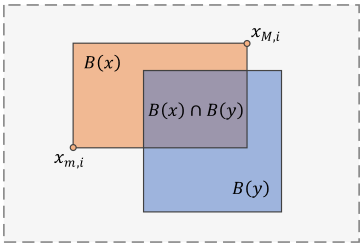



Figure 2: The box embeddings.

CCL 2024

be interpreted as a probability distribution of binary random variables. In this case, the marginal probabilities of the objects are represented as the volume of the corresponding box:

$$V\left[B\left(x\right)\right] = \prod_{i=1}^{d}\left(x_{M,i} - x_{m,i}\right) \tag{1}$$

After that, in order to calculated the relation between different boxes, we introduce the joint and conditional probability. The joint probability of two boxes can be defined as the volume of intersection, since intersection process between two boxes will yield another box. The joint probability can be defined as:

$$V[B(x) \cap B(y)] = \prod_{i=1}^{d} max(min(x_{M,i}, y_{M,i}) - max(x_{m,i}, y_{m,i}), 0) \tag{2}$$

The conditional probabilities between two boxes can be further calculated with marginal probabilities and joint probabilities:

$$P\left(x|y\right) = \frac{V\left[B\left(x\right) \cap B\left(y\right)\right]}{V\left[B\left(y\right)\right]} \tag{3}$$

**Optimization.** The basic box embeddings model suffers from gradient loss during gradient descent-based optimization, which causes the model to fail to learn the embedding representation of objects. The Gumbel distribution is commonly used in probability theory. It models the distribution of the maximum or minimum values of the sample. According to Dasgupta et al. (Dasgupta et al., 2020), they demonstrated the the minimum and maximum values of the box can be represented as MaxGumbel distribution and the MinGumbel distribution with variance $\beta$, respectively, where $\beta$ is the location parameters.

For the two boxes $B\left(x\right)$ and $B\left(y\right)$ represented in $\Omega_H$ using Gumbel box embeddings, the intersection box $B\left(z\right)$ can also be represented by the Gumbel distribution, where the two vector parameters $z_m$ and $z_M$ of $B\left(z\right)$ are calculated as:

$$\boldsymbol{z}_m = \beta \ln\left(e^{\frac{x_m}{\beta}} + e^{\frac{y_m}{\beta}}\right), \boldsymbol{z}_M = -\beta \ln\left(e^{-\frac{x_M}{\beta}} + e^{-\frac{y_M}{\beta}}\right) \tag{4}$$

The volume $V\left[B\left(x\right)\right]$ of the box $B(x)$ based on the Gumbel distribution representation can be approximated as a softplus function. The approximated box volume is calculated as follows.

$$V\left[B\left(x\right)\right] == \prod_{i=1}^{d} softplus\left(\frac{x_{M,i} - x_{m,i}}{\beta} - 2\gamma\right) \tag{5}$$

### 4.3 Multi-task Embedding Attention Unit

Items in the recommendation system usually correspond to many entities. The knowledge graph enhanced recommendation method will establish interaction relations between the items and the corresponding entities. The interaction relations will be established by interaction units. The multi-task embedding attention unit introduces the attention mechanism in the interaction process between items and entities to effectively share the high order interaction feature. The structure of this unit is shown in Figure 3.

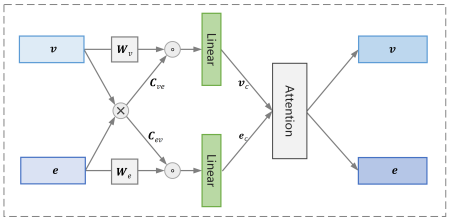



Figure 3: Multi-task embedding attention unit.

The semantic hierarchical structure of word can be modeled by conditional probabilities of box embeddings. For example, there is an edge from "Dog" to "Mammal" in the WordNet. We can convert this edge into the conditional probability of a binary random variable $P\left(Mammal|Dog\right) = 1$, then get the box embeddings of each word and their relationship, as shown in Figure 4.

This unit input item embeddings $\boldsymbol{v} \in \mathbb{R}^d$ and entity embeddings $\boldsymbol{e} \in \mathbb{R}^d$, performs a matrix product operation to construct the feature matrix $\boldsymbol{C}_{ve}$ and $\boldsymbol{C}_{ev}$ between the items and entities:

$$\boldsymbol{C}_{ve} = \boldsymbol{v}^\top \otimes \boldsymbol{e},\ \boldsymbol{C}_{ev} = \boldsymbol{e}^\top \otimes \boldsymbol{v} \tag{6}$$

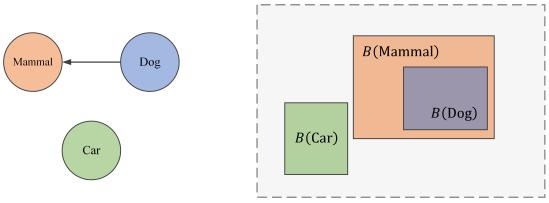



Figure 4: The box embeddings model semantic hierarchical structure.

where $\boldsymbol{C}_{ve}$, $\boldsymbol{C}_{ev} \in \mathbb{R}^{d\times d}$ can explicitly model all interaction features between items and entities. However, the matrix product calculation made features between item and entity over fused, which result in reducing feature weight of the item in the recommendation task. Thus, the feature representations of items and entities are further enriched by four parameter matrices $\boldsymbol{W}_v$, $\boldsymbol{W}_{ve}$, $\boldsymbol{W}_e$, $\boldsymbol{W}_{ev}$. Then preliminary interative feature $\boldsymbol{v}_c$ and $\boldsymbol{e}_c \in \mathbb{R}^d$ of item embeddings and entity embeddings can by obtained by a linear layer:

$$\boldsymbol{v}_c = l\left(\boldsymbol{W}_v\left(\boldsymbol{v}^\top \boldsymbol{v}\right) \circ \boldsymbol{W}_{ve}\boldsymbol{C}_{ve}\right),\ \boldsymbol{e}_c = l\left(\boldsymbol{W}_e\left(\boldsymbol{e}^\top \boldsymbol{e}\right) \circ \boldsymbol{W}_{ev}\boldsymbol{C}_{ev}\right) \tag{7}$$

Where $\circ$ denotes Hadamard product operation, $\boldsymbol{W}_v$, $\boldsymbol{W}_e \in \mathbb{R}^{d\times d}$ are the parameter matrix of item embeddings $\boldsymbol{v}$ and entity embeddings $\boldsymbol{e}$, $\boldsymbol{W}_{ve}$, $\boldsymbol{W}_{ev} \in \mathbb{R}^{d\times d}$ are the parameter matrix of feature matrix$\boldsymbol{C}_{ve}$ and $\boldsymbol{C}_{ev}$, and $l(\cdot)$ denote linear neural network layer.

To further learn the feature weights of item and entity embeddings, the attention mechanism is to model the higher-order interaction features based on the preliminary interaction feature representations. Since attention calculations are involved between the two tasks, it is necessary to define the query $\boldsymbol{q}$ and the key $\boldsymbol{k}$ for item embeddings and entity embeddings, respectively:

$$\boldsymbol{q}_v = \boldsymbol{W}_q\boldsymbol{v}_c\,, \boldsymbol{k}_v = \boldsymbol{W}_k\boldsymbol{v}_c\,, \boldsymbol{q}_e = \boldsymbol{W}_q\boldsymbol{e}_c\,, \boldsymbol{k}_e = \boldsymbol{W}_k\boldsymbol{e}_c \tag{8}$$

Where $\boldsymbol{W}_q$ , $\boldsymbol{W}_k \in \mathbb{R}^{d\times d}$ are the parameters of the query and key. Then the correlation scores $\boldsymbol{s}$ between them are calculated by scaling the dot product. Finally, the correlation scores $\boldsymbol{s}$ are normalized using the softmax function to obtain the attention weight distribution $\alpha$:

$$\alpha_{ve} = softmax\left(\boldsymbol{s}_{ve}\right) = \frac{exp\left(\boldsymbol{s}_{ve}\right)}{\sum_{v,e} exp\left(\boldsymbol{s}_{ve}\right)}\,, \boldsymbol{s}_{ve} = \frac{\boldsymbol{q}_v{}^\top \boldsymbol{k}_e}{\sqrt{d}} \tag{9}$$

The final item $\boldsymbol{v}$ and entity embedding $\boldsymbol{e}$ are obtained from the following equations:

$$\boldsymbol{v} = \alpha_{ve}\boldsymbol{v}_c, \boldsymbol{e} = \alpha_{ev}\boldsymbol{e}_c \tag{10}$$

where $\alpha_{ev}$ is calculated in the same way with $\alpha_{ve}$. After learning from multi-task embedding attention units, the item embedding and entity embedding can effectively capture the higher-order interaction features between items and entities. After the interaction, the learned item embedding and entity embedding are more suitable for the process of their individual task.

### 4.4 Recommendation Module

The recommendation module is used to learn the box embeddings representation of users and items. The module is divided into two parts: the low-level encoding and the box embeddings. Its inputs are the initial embeddings $\boldsymbol{u}$ of users $u$ and $\boldsymbol{v}$ of the items $v$. The user embeddings $\boldsymbol{u}$ are first encoded by the feedforward neural network(FNN) with layer number of $L$ . The encoded user embeddings are represented as $\boldsymbol{u}_L$. The item embeddings $\boldsymbol{v}$ are encoded with the corresponding entity embeddings $\boldsymbol{e}$ in the knowledge graph by L-layer multi-task embedding attention unit (MEAU), and the encoded item embeddings are represented as $\boldsymbol{v}_L$.

$$\boldsymbol{u}_L = FNN\left(\cdots FNN\left(\boldsymbol{u}\right)\right) \tag{11}$$

CCL 2024

$$\boldsymbol{v}_L = \mathbb{E}_{v\sim e}\left(MEAU\left(\cdots MEAU\left(\boldsymbol{v},\boldsymbol{e}\right)[\boldsymbol{v}]\right)\right) \tag{12}$$

where $\mathbb{E}_{v\sim e}$ represents the mapping process between item $v$ and entity $e$. After the low-level encoding, the user and item need to be modeled as box embeddings. For the box, its geometric transformation is mainly translation and scale. Therefore, we further convert the user and item embedding into two vectors. They are the center vector $\boldsymbol{c} \in \mathbb{R}^d$ and the bias vector $\boldsymbol{o} \in \mathbb{R}^d$ . The center vector controls the position of the box, while the bias vector controlls the size of the box. Therefore, the user box $B(u)$ and the item box $B(v)$ are represented as:

$$B\left(u\right) = [(\boldsymbol{u}_m, \boldsymbol{u}_M)] = [\sigma\left(\boldsymbol{c}_u - softplus\left(\boldsymbol{o}_u\right)\right), \sigma\left(\boldsymbol{c}_u + softplus\left(\boldsymbol{o}_u\right)\right)] \tag{13}$$

$$B\left(v\right) = [(\boldsymbol{v}_m, \boldsymbol{v}_M)] = [\sigma\left(\boldsymbol{c}_v - softplus\left(\boldsymbol{o}_v\right)\right), \sigma\left(\boldsymbol{c}_v + softplus\left(\boldsymbol{o}_v\right)\right)] \tag{14}$$

Finally, based on the conditional probability of the box, the probability of user $u$ is interested in the item $v$ is predicted as $\hat{y}_{uv}$:

$$\hat{y}_{uv} = P\left(\ B\left(v\right)|B\left(u\right)\right) = \frac{V\left[B\left(v\right) \cap B\left(u\right)\right]}{V\left(B\left(u\right)\right)} \tag{15}$$

### 4.5 Knowledge Graph Embedding Module

The knowledge graph embedding module is used to learn the box embeddings representation of entities and relations in the knowledge graph embedding task. Its network structure is similar to the recommendation module. Its inputs are the initial feature embeddings $\boldsymbol{h}$ of head entities $h$, $\boldsymbol{r}$ of relations $r$, and $\boldsymbol{t}$ of tail entities $t$ in the knowledge graph triples $(h, r, t)$. The relation embedding $\boldsymbol{r}$ is first encoded using an L-layer FNN, and the encoded relation embeddings are represented as $\boldsymbol{r}_L$. The head entity embeddings $\boldsymbol{h}$ are encoded with the corresponding item embeddings $\boldsymbol{v}$ in the recommendation by the L-layer MEAU, and the encoded head entity embeddings are represented as $\boldsymbol{h}_L$.

$$\boldsymbol{r}_L = FNN\left(\cdots FNN\left(\boldsymbol{r}\right)\right) \tag{16}$$

$$\boldsymbol{h}_L = \mathbb{E}_{h\sim v}\left(MEAU\left(\cdots MEAU\left(\boldsymbol{h},\boldsymbol{v}\right)[\boldsymbol{h}]\right)\right) \tag{17}$$

where $\mathbb{E}_{h\sim v}$ represents the mapping process between the head entity $h$ and the item $v$. After completing the lower-level encoding of relations and entities, the box embeddings of relations and entities need to be modeled. Where the box representation of entities is similar to the box representation of users and items in the recommendation module. But in the knowledge graph, multiple types of relationships often exist between entities, the embedding of the same entity under different relationships should be different. Therefore, we further represent the relation $r$ between entities as two transformation vectors, which are the center transformation vector $\boldsymbol{c}^r \in \mathbb{R}^d$ and the bias transformation vector $\boldsymbol{o}^r \in \mathbb{R}^d$ controlling the position and size of the entity box, respectively, and the transformation vectors of different relations can transform the position and size of entity box in different ways. For the head entity box $B(h)$ and the tail entity box $B(t)$, their center vectors and bias vectors under the relation $r$ are represented as follows:

$$\boldsymbol{c}_h\left(r\right) = \boldsymbol{c}_h + \boldsymbol{c}_h^r\ ,\ \boldsymbol{o}_h\left(r\right) = \boldsymbol{o}_h \circ \boldsymbol{o}_h^r \tag{18}$$

$$\boldsymbol{c}_t\left(r\right) = \boldsymbol{c}_t + \boldsymbol{c}_t^r\ ,\ \boldsymbol{o}_t\left(r\right) = \boldsymbol{o}_t \circ \boldsymbol{o}_t^r \tag{19}$$

where $\circ$ is the Hadamard product, $\boldsymbol{c}_h\left(r\right),\ \boldsymbol{o}_h\left(r\right)$ and $\boldsymbol{c}_t\left(r\right),\ \boldsymbol{o}_t\left(r\right)$ represent the center vector and bias vector of the head entity box $B(h)$ and the tail entity box $B(t)$ after the transformation of the relation $r$, respectively. Based on the above center vectors and bias vectors, the box embeddings of $B(h)$ and $B(t)$ under the relation $r$ are given as $T_{rh}\left(B\left(h\right)\right)$ and $T_{rt}\left(B\left(t\right)\right)$.

$$T_{rh}\left(B\left(h\right)\right) = [\sigma\left(\boldsymbol{c}_h\left(r\right) - softplus\left(\boldsymbol{o}_h\left(r\right)\right)\right), \sigma\left(\boldsymbol{c}_h\left(r\right) + softplus\left(\boldsymbol{o}_h\left(r\right)\right)\right)\ ] \tag{20}$$

$$T_{rt}\left(B\left(t\right)\right) = [\sigma\left(\boldsymbol{c}_t\left(r\right) - softplus\left(\boldsymbol{o}_t\left(r\right)\right)\right), \sigma\left(\boldsymbol{c}_t\left(r\right) + softplus\left(\boldsymbol{o}_t\left(r\right)\right)\right)\ ] \tag{21}$$

Finally, based on the geometric properties of the box, the score $f_r\left(h,t\right)$ of the triple $(h, r, t)$ is predicted:

$$f_r\left(h,t\right) = \frac{V\left[T_{rh}\left(B\left(h\right)\right) \cap T_{rt}\left(B\left(t\right)\right)\right]}{V\left[T_{rt}\left(B\left(t\right)\right)\right]} \tag{22}$$

CCL 2024

### 4.6 Learning

The BoxMAKR model is jointly trained by the recommendation task and the knowledge graph embedding task. The learning objective $\mathcal{L}$ includes three parts: the recommendation task loss function $\mathcal{L}_{REC}$, the knowledge graph embedding task loss function $\mathcal{L}_{KGC}$, and the regular term $\mathcal{L}_2$. The equation is as follows:

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_{REC} + \mathcal{L}_{KGC} + \mathcal{L}_2 \\ &= \sum_{u \in U, v \in V} g\left(\hat{y}_{uv}, y_{uv}\right) - \lambda_1 \left( \sum_{(h,r,t) \in \mathcal{G}} f_r\left(h,t\right) - \sum_{(h',r,t') \notin S} f_r\left(h',t'\right) \right) + \lambda_2 \left\| \boldsymbol{W} \right\|_2^2 \end{aligned} \tag{23}$$

where $g\left(\cdot\right)$ is the cross-entropy loss function, the learning objective of the recommendation task is to make the prediction probability $\hat{y}_{uv}$ close to the real probability $y_{uv}$. In Equation 23, $(h', r, t')$ is the negative sample of the triples. The learning objective of the knowledge graph embedding task is to improve the prediction score $f_r\left(h,t\right)$ , while reducing the prediction score $f_r\left(h',t'\right)$ of the negative samples. Finally, the $L_2$ regular term is to avoid model overfitting. $lambda_1$ and $lambda_2$ are the hyperparameters used to balance the knowledge graph embedding task and the regular term.

| Datasets | Users | Items | Interactions | KG triples | Sparsity |
|---|---|---|---|---|---|
| MovieLens 1M | 6036 | 2347 | 753772 | 20195 | 94.68% |
| Book-Crossing | 17860 | 14910 | 139746 | 19793 | 99.95% |
| Last.FM | 1872 | 3846 | 42346 | 15518 | 99.41% |

Table 1: Basic statistics for the three datasets.

## 5 Experiments

### 5.1 Datasets

In our experiments, we used three publicly available datasets containing explicit user feedback. MovieLens-1M is a movie recommendation benchmark dataset which described the users' preferences for movies. It includes 6036 users for 2347 movies with 753,772 ratings. Its data sparsity is 94.68%. Book-Crossing is crawled from the Book-Crossing community, which described users' grades for book. In total, it includes 139,746 ratings from 17,860 users on 14,910 book items. Its data sparsity is 99.95%. Last.FM is from the Last.FM online music system, which described the listen records of users and music. It includes a total number of 42,346 listening counts, which produced by 1,872 users about 3,846 songs. Its data sparsity is 99.41%.

We divided the users feedback into positive and negative. The positive feedback of users to items is denoted as "1", while negative feedback is denoted as "0". In MovieLens-1M, user ratings of 4 and 5 were selected as positive feedback. However, the other two datasets have a large data sparsity. We use all user interactions on items in the dataset as positive feedback. After building the positive feedback, we randomly sampled the same number of negative samples for each user with items that are not interaction. Then, we further used the Microsoft Satori service to build the knowledge graph on the above datasets. The statistics of the three datasets are shown in Table 1.

### 5.2 Baselines

To verify the validity of our model, we compare it with six classical baseline models. KTUP (Cao et al., 2019) jointly learned the recommendation task and the knowledge graph complementation task. The core idea of this method is introducing preference induction to model users' preference features for items. The bridge between the two tasks is the alignment of the recommended items with the corresponding entities. The user preferences are related to the relations between the entities. RippleNet (Wang et al., 2018) introduced the concept of preference propagation to optimize user features by sampling from the knowledge graph in an aggregated way. Eventually, user preferences from historical interaction information are propagated over paths in knowledge graph to realize the joint training of knowledge graph and

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1172-1182, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

| Datasets | Dim | Layer | Lr1 | Lr2 | Lr3 |
|---|---|---|---|---|---|
| MovieLens-1M | 32 | 4 | 3e-3 | 1e-3 | 0.5 |
| Book-Crossing | 16 | 3 | 5e-4 | 2e-5 | 0.1 |
| Last.FM | 16 | 2 | 3e-3 | 2e-4 | 0.1 |

Table 2: Hyperparameters settings for the three datasets.

recommendation. MKR (Wang et al., 2019) is an unified framework for knowledge graph enhanced recommendation with multi-task learning. The framework includes recommendation and knowledge graph embedding module, which are connected by cross&compress unit. During training, they used alternative learning strategy to train recommendation and knowledge graph embedding task. Ripp-MKR (Wang et al., 2021) imporved MKR model with RippleNet. In their work, Ripp-MKR combined the knowledge graph with the user history interaction information to represent the user's features. TransMKR (Hu et al., 2022) improved the knowledge graph embedding module of MKR by using the TransR algorithm. TransR can enhance the representation capability of embedding knowledge graph.

### 5.3 Experiments setup

In the experiments, the hyperparameters of the baseline models are initialized using the data from the original work, and the optimal parameter is selected by fine-tuning in our experiments. The hyperparameter settings of BoxMAKR model on different datasets are shown in the table 2, where $dim$ represents the embedding dimension, $layer$ represents the number of network layers, $lr_1$ and $lr_2$ are the learning rates of the recommendation task and the knowledge graph embedding task, respectively. $lambda_1$ is the weight parameter of the recommendation task. The other hyperparameters are fixed in all datasets. The weight $\lambda_2$ of the regular term is 1e-6. The scale parameter $\beta$ of the Gumbel distribution in the box embeddings is 0.01. The training interval $t$ of the knowledge graph embedding task is 3.

The datasets are divided into the training, validation and test set with the ratio of 6:2:2. Each experiment is conducted three times, and the average results are reported to validate the robustness of the model. We compared all methods in two typical task. Click-through rate (CTR) prediction. The trained model is applied to the test set to get click-through rate prediction scores, and the performance is evaluated using two evaluation metrics, $AUC$ and $ACC$. TOP-K recommendation. The trained model is applied to the test set, and the K recommendation items with the highest predicted click probability are selected for each user. The performance is evaluated using two evaluation metrics, $Recall@K$ and $F1@K$.

| Model | MovieLens-1M | | Book-Crossing | | Last.FM | |
|---|---|---|---|---|---|---|
| | AUC | ACC | AUC | ACC | AUC | ACC |
| KTUP | 0.822 | 0.761 | 0.689 | 0.643 | 0.735 | 0.667 |
| RippleNet | 0.920 | 0.842 | 0.729 | 0.662 | 0.768 | 0.691 |
| MKR | 0.917 | 0.843 | 0.734 | 0.704 | 0.799 | 0.752 |
| Ripp-MKR | 0.922 | 0.845 | 0.740 | 0.712 | 0.799 | **0.756** |
| TransMKR | 0.916 | 0.838 | 0.737 | 0.706 | 0.797 | 0.753 |
| **BoxMAKR(ours)** | **0.926** | **0.855** | **0.749** | **0.712** | **0.804** | 0.752 |

Table 3: The results of AUC and Accuracy in CTR prediction.

## 6 Results

### 6.1 Overall results

The experimental results of the CTR prediction and TOP-K recommendation are showed in table 3. The comparison between different models on Top-K in three datasets are shown in Figure 5, Figure 6 and Figure 7, respectively.

From Table 3, we observed that KTUP performs the worst in the three datasets, especially in the dense MovieLens-1M where the performance is much lower than the other models. This demonstrated that

the preference induction introduced by KTUP cannot build effective interactions between users and the knowledge graph which is not suitable for complex recommendation scenarios.

RippleNet has better performance in the MovieLens-1M because the user-item interaction information is dense and preference propagation can more accurately capture user interests. However, RippleNet does not perform as well as MKR in the Book-Crossing and Last.FM when the data is sparse.

In Table 3, Ripp-MKR and TransMKR are used the same framework of MKR. Ripp-MKR combined the advantages of RippleNet and MKR, has better performance on all datasets. It shows that introducing user preferences under a multi-task learning strategy can effectively integrate the knowledge graph embedding and recommendation task. TransMKR has no significant improvement in overall performance compared with MKR. This is because TransMKR only improves the embedding method of the knowledge graph embedding module without adapting the recommendation module and interaction unit. It leads to the recommendation task not being able to effectively obtain the feature representation in the knowledge graph embedding task.

Compared with other baselines, the BoxMAKR improved the $AUC$ scores by 0.9%, 1.5%, and 0.5% on the three datasets, respectively. Meanwhile, the $ACC$ scores improve by 1.2% and 0.8% on the MovieLens-1M and Book-Crossing, respectively. But not on the Last.FM, which is probably because the amount of data is too small to fully train the network model. BoxMAKR outperforms the optimal result on the three datasets.

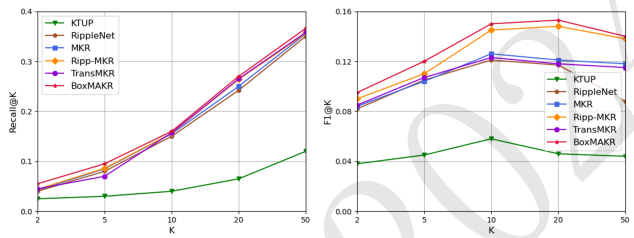



Figure 5: TOP-K recommendation results on the MovieLens-1M.

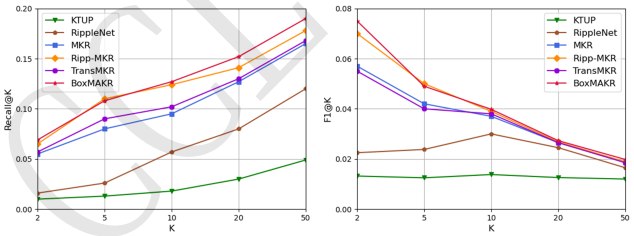



Figure 6: TOP-K recommendation results on the Book-Crossing.

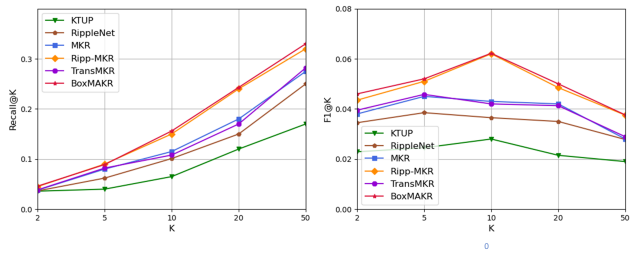



Figure 7: TOP-K recommendation results on the Last.FM.

### 6.2 Ablation

We further demonstrated the effectiveness of box embeddings and multi-task embedding attention unit through ablation experiments.

In Table 4, BoxMAKR(w/o Att) and BoxMAKR(w/o box) are models of BoxMAKR using MKR's interaction unit and prediction network, respectively. Compared with BoxMAKR, BoxMAKR(w/o Att) and BoxMAKR(w/o box) decreased the $AUC$ metrics by 0.5% and 0.7% on average and the $ACC$ metrics by 0.6% and 0.8% on average in the three datasets. The results of the ablation experiments show that introducing attention mechanisms in the interaction units can share the higher-order interaction features between items and entities more effectively. It can also be seen that box embeddings play a very important role in BoxMAKR because box embeddings can provide more semantic information for recommendation tasks and knowledge graph embedding tasks by modeling the semantic hierarchical structural relationships between data.

| Model | MovieLens-1M | | Book-Crossing | | Last.FM | |
|---|---|---|---|---|---|---|
| | AUC | ACC | AUC | ACC | AUC | ACC |
| BoxMAKR(w/o Att) | 0.922 | 0.846 | 0.745 | 0.707 | 0.799 | 0.748 |
| BoxMAKR(w/o Box) | 0.920 | 0.842 | 0.742 | 0.704 | 0.796 | 0.749 |
| **BoxMAKR** | **0.926** | **0.855** | **0.749** | **0.712** | **0.804** | **0.752** |

Table 4: The ablation results of AUC and Accuracy in CTR prediction.

## 7 Conclusions

This paper proposed a box embedding method for knowledge graph enhanced recommendation system. Based on the MKR model, we use box embeddings to represent users, items and entities. This embeddings allow us to model the semantic hierarchical structure relationships between those objects. Further, the multi-task attention method learned the higher-order interactions between items and corresponding entities. We conduct extensive experiments in three recommendation datasets. The results showed that our method is better than the original model MKR, and have great advantages over other baseline models.

CCL 2024

## Acknowledgements

This work is supported by National Natural Science Foundation of China (Nos.62066033, 61966025); Inner Mongolia Natural Science Foundation (Nos.2024MS06013, 2022JQ05); Inner Mongolia Autonomous Region Science and Technology Programme Project (Nos.2023YFSW0001, 2022YFDZ0059); Collaborative Innovation Project of Universities and Institutes in Hohhot City. We also thank all the anonymous reviewers for their insightful comments.

## References

Yixin Cao, Xiang Wang, Xiangnan He, Zikun Hu, and Tat-Seng Chua. 2019. Unifying knowledge graph learning and recommendation: Towards a better understanding of user preferences. In *The World Wide Web Conference*, pages 151–161.

Shib Dasgupta, Michael Boratko, Dongxu Zhang, Luke Vilnis, Xiang Li, and Andrew McCallum. 2020. Improving local identifiability in probabilistic box embeddings. *Advances in Neural Information Processing Systems*, 33:182–192.

Yifan Hu, Yehuda Koren, and Chris Volinsky. 2008. Collaborative filtering for implicit feedback datasets. In *2008 Eighth IEEE international conference on data mining*, pages 263–272. IEEE.

Bojing Hu, Yaqin Ye, Yingqiang Zhong, Jiao Pan, and Maosheng Hu. 2022. TransMKR: Translation-based knowledge graph enhanced multi-task point-of-interest recommendation. *Neurocomputing*, 474:107–114.

Xueyong Jiang, Baisong Liu, Jiangchen Qin, Yunchong Zhang, and Jiangbo Qian. 2022. Fedncf: Federated neural collaborative filtering for privacy-preserving recommender system. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Zahid Younas Khan, Zhendong Niu, Ally S Nyamawe, and Ijaz ul Haq. 2021. A deep hybrid model for recommendation by jointly leveraging ratings, reviews and metadata information. *Engineering Applications of Artificial Intelligence*, 97:104066–104066.

R Kiran, Pradeep Kumar, and Bharat Bhasker. 2020. Dnnrec: A novel deep learning based hybrid recommender system. *Expert Systems with Applications*, 144:113054–113054.

Alice Lai and Julia Hockenmaier. 2017. Learning to predict denotational probabilities for modeling entailment. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 721–730.

Qiuyu Liang, Weihua Wang, Feilong Bao, and Guanglai Gao. 2024. $L^2$GC:lorentzian linear graph convolutional networks for node classification. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9988–9998, Torino, Italia, May. ELRA and ICCL.

Xinyu Liu and Zengmao Wang. 2022. Cfda: Collaborative filtering with dual autoencoder for recommender system. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7.

Nickel Maximillian and Douwe Kiela. 2018. Learning continuous hierarchies in the lorentz model of hyperbolic geometry. In *International Conference on Machine Learning*, pages 3779–3788. PMLR.

Maximillian Nickel and Douwe Kiela. 2017. Poincaré embeddings for learning hierarchical representations. *Advances in Neural Information Processing Systems*, 30.

Yasumasa Onoe, Michael Boratko, Andrew McCallum, and Greg Durrett. 2021. Modeling fine-grained entity types with box embeddings. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*.

Dhruvesh Patel and Shib Sankar. 2020. Representing joint hierarchies with box embeddings. *Automated Knowledge Base Construction*.

Steffen Rendle. 2010. Factorization machines. In *2010 IEEE International conference on data mining*, pages 995–1000. IEEE.

Luke Vilnis, Xiang Li, Shikhar Murty, and Andrew McCallum. 2018. Probabilistic embedding of knowledge graphs with box lattice measures. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 263–272.

Hongwei Wang, Fuzheng Zhang, Jialin Wang, Miao Zhao, Wenjie Li, Xing Xie, and Minyi Guo. 2018. Ripplenet: Propagating user preferences on the knowledge graph for recommender systems. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 417–426.

Hongwei Wang, Fuzheng Zhang, Miao Zhao, Wenjie Li, Xing Xie, and Minyi Guo. 2019. Multi-task feature learning for knowledge graph enhanced recommendation. In *The World Wide Web Conference*, pages 2000–2010.

Chao Wang, Hengshu Zhu, Chen Zhu, Chuan Qin, and Hui Xiong. 2020. Setrank: A setwise bayesian approach for collaborative ranking from implicit feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6127–6136.

Yuequn Wang, Liyan Dong, Yongli Li, and Hao Zhang. 2021. Multitask feature learning approach for knowledge graph enhanced recommendations with ripplenet. *PLoS ONE*, 16.

Sheng Zhang, Weihong Wang, James Ford, and Fillia Makedon. 2006. Learning from incomplete ratings using non-negative matrix factorization. In *Proceedings of the 2006 SIAM International Conference on Data Mining*, pages 549–553. SIAM.

Fuzheng Zhang, Nicholas Jing Yuan, Defu Lian, Xing Xie, and Wei-Ying Ma. 2016. Collaborative knowledge base embedding for recommender systems. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 353–362.

CCL 2024

# Readability-guided Idiom-aware Sentence Simplification (RISS) for Chinese

**Jingshen Zhang[1], Xinglu Chen[1], Xinying Qiu[1]*, Zhimin Wang[2], Wenhe Feng[3]**
[1]School of Information Science and Technology,
[2]Faculty of Chinese Language and Culture
[3]Laboratory of Language Engineering and Computing
Guangdong University of Foreign Studies, Guangzhou, China
audbut0702@163.com, luc88749@gmail.com,
xy.qiu@foxmail.com, wangzm000@qq.com, wenhefeng@gdufs.edu.cn

## Abstract

Chinese sentence simplification faces challenges due to the lack of large-scale labeled parallel corpora and the prevalence of idioms. To address these challenges, we propose Readability-guided Idiom-aware Sentence Simplification (RISS), a novel framework that combines data augmentation techniques. RISS introduces two key components: (1) Readability-guided Paraphrase Selection (RPS), a method for mining high-quality sentence pairs, and (2) Idiom-aware Simplification (IAS), a model that enhances the comprehension and simplification of idiomatic expressions. By integrating RPS and IAS using multi-stage and multi-task learning strategies, RISS outperforms previous state-of-the-art methods on two Chinese sentence simplification datasets. Furthermore, RISS achieves additional improvements when fine-tuned on a small labeled dataset. Our approach demonstrates the potential for more effective and accessible Chinese text simplification.

## 1 Introduction

Sentence simplification (SS) aims to reduce the linguistic complexity of a sentence while preserving its meaning (Al-Thanyyan and Azmi, 2021). It has various societal applications, such as increasing accessibility for people with cognitive disabilities (Stajner, 2021) and second language learners (Petersen and Ostendorf, 2007). While English sentence simplification has made significant progress due to the availability of large-scale parallel corpora, other languages face the challenge of limited publicly available resources (Ryan et al., 2023). For Chinese, in particular, the publicly accessible sentence simplification corpora contain only around 5,000 entries (Sun et al., 2023; Chong et al., 2023). This scarcity of labeled data is a primary hurdle for advancing Chinese sentence simplification research.

In addition to the lack of parallel corpora, Chinese sentence simplification faces another challenge: the prevalence of idioms. Idioms are non-compositional expressions whose metaphorical meanings deviate from the literal meanings of their constituent words (Bobrow and Bell, 1973; Li et al., 2023). In Chinese, idioms are frequently used in sentences, and their simplification often requires providing a brief explanatory sentence rather than simply replacing them with synonyms, as is common in traditional lexical simplification.

To address these challenges, We propose Readability-guided Idiom-aware Sentence Simplification (RISS), a novel framework combining data augmentation and transfer learning to address the challenges in Chinese sentence simplification. RISS integrates Readability-guided Paraphrase Selection (RPS), which mines high-quality sentence pairs from Chinese paraphrases, and Idiom-aware Simplification (IAS), an idiom-based transfer learning strategy. Through multi-stage and multi-task learning, RISS outperforms state-of-the-art methods on two datasets without labeled data and further improves when fine-tuned on a small labeled dataset.

The paper is organized as follows: Section 2 reviews related work, Section 3 describes

* Corresponding author
©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1183-1200, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

our methodologies (RPS, IAS, and RISS), Section 4 outlines the experimental setup, Section 5 presents and analyzes the results, and Section 6 provides conclusions and future directions. The code, data and outputs of the proposed method are available at https://github.com/astro-jon/RISS

## 2 Related Work

### 2.1 Data Augmentation in Sentence Simplification

Data augmentation techniques play a crucial role in addressing the scarcity of training data for sentence simplification, especially for languages other than English. Various approaches have been proposed to create pseudo-parallel corpora, such as mining paraphrases from large-scale datasets (Zhao et al., 2020; Martin et al., 2022), utilizing back-translation (Lu et al., 2021), leveraging summarization datasets (Sun et al., 2023), and using Wikipedia and Simple Wikipedia articles (Kriz et al., 2020). Chi et al.(2023) focused on creating a dataset with sentences at different complexity levels using a paraphrase corpus labeled with difficulty scores.

However, the quality of the augmented data is essential for effectively training simplification models. Factual errors, such as inadequate simplifications or misalignments, can negatively impact model performance (Xu et al., 2015). To mitigate this issue, researchers have proposed methods to identify and down-weight sentence pairs with factual errors during training (Ma et al., 2022), combine rule-based sentence splitting and deletion with paraphrasing (Jiang et al., 2020), and leverage sentence alignment, paraphrasing, and simplification-specific noise reduction techniques (Maddela et al., 2021).

Despite these advancements, there is still a need for methods that can effectively select high-quality paraphrase pairs that are most suitable for the simplification task. Our proposed Readability-guided Paraphrase Selection (RPS) framework addresses this gap by automatically selecting high-quality paraphrase pairs based on readability differences and linguistic similarities.

### 2.2 Lexical Simplification

Lexical simplification is a key component of sentence simplification that aims to replace complex words or phrases with simpler alternatives while preserving the original meaning (Shardlow, 2014). This process is essential for improving the readability and accessibility of texts for various audiences (Paetzold and Specia, 2017). Previous works have explored different approaches to lexical simplification, such as learning explicit editing operations (Dong et al., 2019), neural ranking (Paetzold and Specia, 2017), and leveraging word-complexity lexicons and neural readability ranking models (Maddela and Xu, 2018).

However, lexical simplification faces challenges in handling idioms and non-compositional expressions (Abrahamsson et al., 2014; Rello et al., 2013). Recent research has explored integrating idiomatic knowledge into sentence simplification models (Yimam et al., 2018) and proposed multilingual lexical simplification methods via paraphrase generation (Qiang et al., 2023). Wan et al.(2023) introduced a strategy that dynamically identifies and masks complex words in complicated sentences and simple words in simple sentences to enhance lexical simplification.

Despite these advancements, the simplification of idioms remains a significant challenge due to their figurative and context-dependent meanings (Rello et al., 2013). Our proposed Idiom-aware Simplification (IAS) module addresses this challenge by leveraging idiom paraphrasing to enhance the readability and comprehensibility of simplified text, aligning with the broader goal of making written content more accessible for all readers (Shardlow, 2014).

## 3 Methodology

Our proposed Readability-guided Idiom-aware Sentence Simplification framework combines two novel methodologies: RPS for mining high-quality paraphrase data and IAS for idiomatic transfer learning. We integrate these modules using multi-stage and multi-task learning strate-

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1183-1200, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

gies to enhance the performance of sentence simplification models. We provide a glossary of abbreviations in Appendix A.

### 3.1 RPS: Readability-guided Paraphrase Selection

The RPS module, as shown in Figure 1, automatically mines high-quality paraphrase sentence pairs to serve as training data for sentence simplification models. Given a set of paraphrase data, inspired by previous research(Bandel et al., 2022), we calculate pairwise readability differences, Levenshtein similarities, syntactic similarities, and semantic similarities for each sentence pair. These features are then integrated into a quality scoring algorithm that selects the highest-quality(Hi-Q) paraphrase pairs based on predefined thresholds.

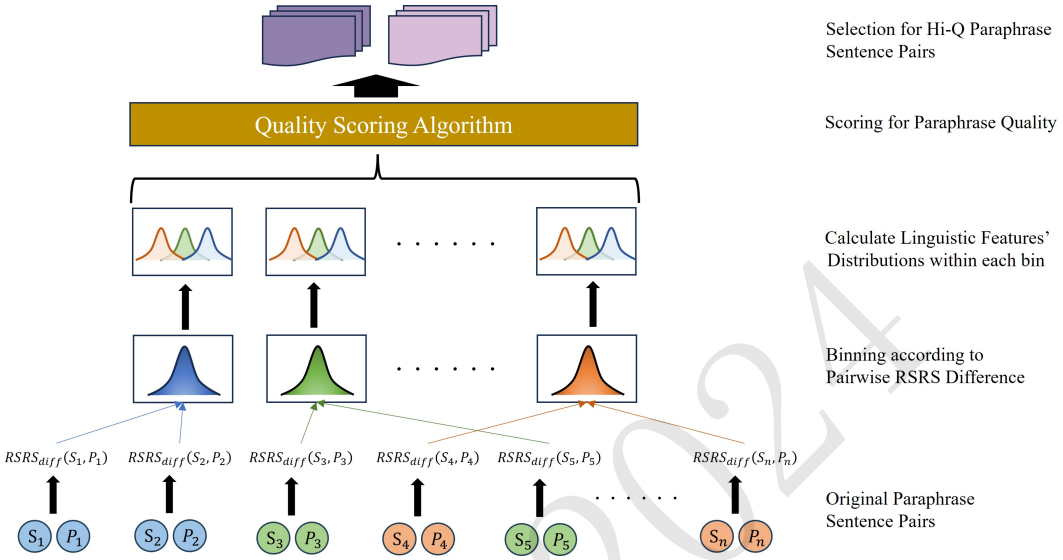



Figure 1: The architecture of the proposed RPS framework

**1. Pairwise Readability Difference:** Given a paraphrase sentence pair $(S_i, P_i)$, we compute their readability scores, $RSRS_i$ and $RSRS_j$, using the $RSRS$ (Martinc et al., 2021), a model evaluation-based readability metric. The RSRS score is calculated as follow:

$$RSRS = \frac{\sum_{i=1}^{S} \sqrt{i} \cdot WNLL(i)}{S} \tag{1}$$

where $S$ represents the length of the sentence, $i$ represents the rank of a word in this sentence (sorted in ascending order by $WNLL$ value), and $WNLL$ is the word-level negative log-likelihood of the $i-th$ word, computed as:

$$WNLL = -(y_t log y_p + (1-y_t) log(1-y_p)) \tag{2}$$

where $y_p$ denotes the probability predicted by the language model based on the historical sequence, and $y_t$ denotes the empirical distribution for a specific position in the sentence, (i.e., $y_t = 1$ for the actual next word and 0 for all the other words in the vocabulary).

We choose distilbert-base-multilingual-cased(Sanh et al., 2019) as the RSRS discriminant model, which has been experimentally proven effective (see Appendix E for details). The pairwise readability difference, defined as $RSRS_{Diff}(S_i, P_i) = RSRS(S_i) - RSRS(P_i)$, quantifies the simplification difference between the original sentence and its paraphrase. Considering this difference allows selecting paraphrase pairs with desirable simplification characteristics to enhance sentence simplification model training.

CCL 2024

**2. Pairwise Levenshtein Similarity:** We calculate the Levenshtein similarity between a paraphrase sentence pair $(S_i, P_i)$ as follow:

$$\text{Lev}_{Sim}(S_i, P_i) = \frac{sum(S_i, P_i) - ldist(S_i, P_i)}{sum(S_i, P_i)} \tag{3}$$

where the $sum(S_i, P_i)$ is the total length of $S_i$ and $P_i$, $ldist(S_i, P_i)$ is the weighted edit distance between $S_i$ and $P_i$ based on Levenshtein distance (Levenshtein, 1966). The weighted distance is calculated as :

$$ldist(S_i, P_i) = Num(INSERT) + Num(DELETE) + 2 * Num(REPLACE) \tag{4}$$

where $Num(\cdot)$ denotes the count of the respective edit operation. The motivation behind using Levenshtein similarity is to ensure that the selected paraphrase pairs maintain a sufficient level of lexical similarity, which is important for preserving the core meaning of the original sentence during simplification.

**3. Pairwise Syntactic Similarity:** To measure the syntactic similarity between a paraphrase sentence pair $(S_i, P_i)$, we first construct the third-level constituency parse trees for both sentences. We then calculate the syntactic similarity as follows:

$$\text{Syn}_{Sim}(S_i, P_i) = 1 - \frac{TreeEditDistance(\text{Parse}_{Tree}(S_i), \text{Parse}_{Tree}(P_i))}{\max(\text{len}(\text{Parse}_{Tree}(S_i)), \text{len}(\text{Parse}_{Tree}(P_i)))} \tag{5}$$

where $\text{Parse}_{Tree}(S_i)$ and $\text{Parse}_{Tree}(P_i)$ represent the third-level constituency parse trees of $S_i$ and $P_i$, respectively, and $TreeEditDistance(\cdot, \cdot)$ represents normalized tree edit distance (Zhang and Shasha, 1989). The tree edit distance computes the minimum number of edit operations required to transform one parse tree into the other. Using syntactic similarity helps select paraphrase pairs with similar structures, preserving grammatical integrity in simplified sentences. Pairs with high syntactic similarity provide consistent patterns for learning simplification transformations, making them effective for training sentence simplification models.

**4. Pairwise Semantic Similarity:** To measure the semantic similarity between a paraphrase sentence pair $(S_i, P_i)$, we employ the text2vec[0] embedding approach. We extract the vector representations of both sentences using text2vec model[1] and calculate their cosine similarity as follows:

$$\text{Sem}_{Sim}(S_i, P_i) = \frac{\sum_{j=1}^{n}(text2vec(S_i)_j \times text2vec(P_i)_j)}{\sqrt{\sum_{j=1}^{n} text2vec(S_i)_j^2} \times \sqrt{\sum_{j=1}^{n} text2vec(P_i)_j^2}} \tag{6}$$

where $\text{text2vec}(S_i)$ and $\text{text2vec}(P_i)$ denote the vector representations of $S_i$ and $P_i$, respectively, obtained from the text2vec model, $n$ denotes the dim of embedding.

Using semantic similarity ensures that selected paraphrase pairs preserve the original sentence's core meaning and information during simplification.

**Correlation among features:** Figure 2 shows that higher readability differences correspond to lower linguistic similarity scores in paraphrase pairs. Pearson correlation coefficients confirm strong, statistically significant positive correlations between readability difference and Levenshtein ($r = 0.653$), syntactic ($r = 0.612$), and semantic ($r = 0.624$) similarities ($p < 0.05$). The imperfect correlations ($r < 1$) suggest each feature captures distinct aspects of the original-simplified sentence relationship. These correlations validate our approach of using these features with readability differences to select high-quality paraphrase pairs for sentence simplification task.

[0] https://pypi.org/project/text2vec/

[1] https://huggingface.co/shibing624/text2vec-base-chinese

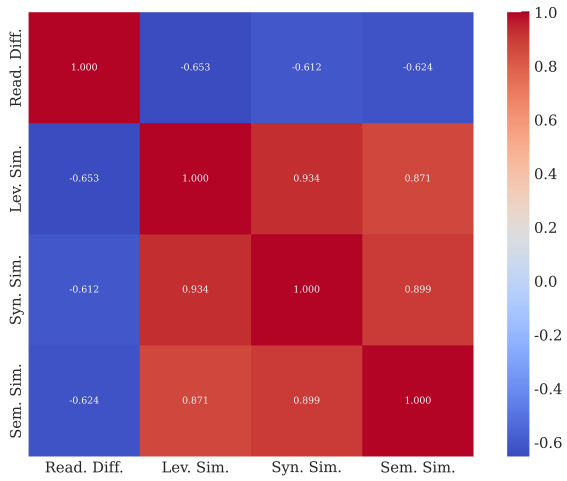



Figure 2: The heat map displays the correlations between readability differences and lexical, syntactic, and semantic similarities. Cool tones indicate that higher readability differences correlate with lower feature values.

**Readability-guided Paraphrase Selection (RPS) Module:** The RPS framework (Figure 1) mines high-quality paraphrase pairs for training sentence simplification models through the following steps:

1. Readability assessment and binning of the paraphrase corpus using the $RSRS_{Diff}$ scores.

2. Extraction of linguistic features ($Lev_{Sim}$, $Syn_{Sim}$, $Sem_{Sim}$) for each pair within each bin.

3. Quality scoring (Eq. 7) to select the best pairs from each bin based on feature means and standard deviations.

$$Q(p,b) = \begin{cases} 1, & \forall \\ & f \in \{lex, syn, sem\} \end{cases} (\mu_f^b - \sigma_f^b \leq f(p) \leq \mu_f^b + \sigma_f^b) \\ 0, \quad \text{otherwise} \end{cases} \tag{7}$$

   where $p$ represents a paraphrase sentence pair, $b$ represents a bin, $f$ represents a feature ($\text{Lex}_{Sim}$, $\text{Syn}_{Sim}$, or $\text{Sem}_{Sim}$), $\mu_f^b$ represents the mean of feature $f$ in bin $b$, $\sigma_f^b$ represents the standard deviation of feature $f$ in bin $b$, $f(p)$ represents the scoring value of feature $f$ for pair $p$.

4. Creation of a curated dataset of high-quality pairs for training simplification models. By considering multiple linguistic dimensions and their relationships with readability differences, RPS identifies paraphrase pairs well-suited for simplification while balancing simplification and fidelity to the original meaning.

### 3.2 Idiom-aware Simplificaiton (IAS) Module

As shown Appendix C, we observe a fair amount of idiom prevalance in the Chinsese Sentence Simplification Datasets. This observation motivates us to propose an idom-aware simplification approach. The IAS module (Figure 3) enhances the sentence simplification model's ability to understand and simplify idioms. IAS introduces an idiom-specific loss component to encourage accurate representations and simplifications of idiomatic phrases. During training, IAS augments the standard sentence-level loss (Eq. 8) with an idiomatic loss term (Eq. 9) that focuses on generating correct idiom simplifications. By jointly optimizing both loss terms (Eq. 10)., IAS guides the model to learn general simplification patterns and specific idiom

CCL 2024

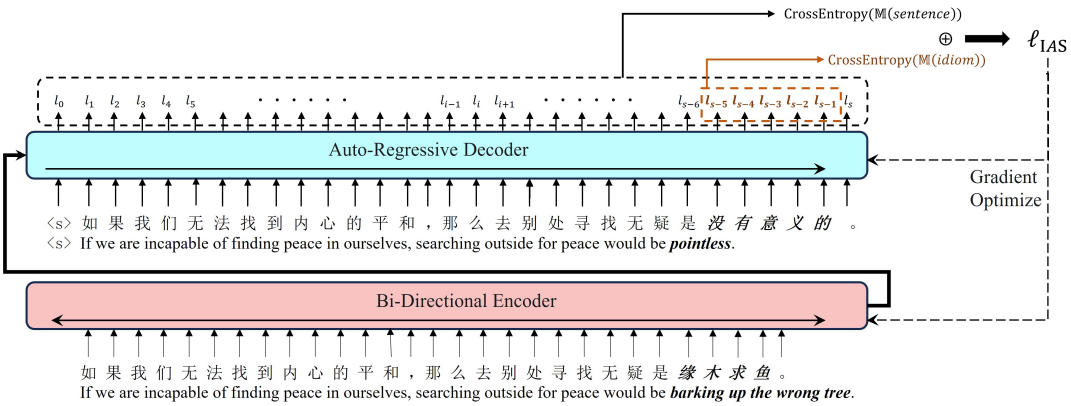



Figure 3: The IAS module architecture. $l_i$ denotes the logits predicted by model $\mathbb{M}$ for the labeled sentence. Bolded italics show the idiom and its explanation.

simplification rules, helping it produce more coherent and understandable simplified sentences.

$$\ell_{\text{sentence}} = \text{CrossEntropy}(\mathbb{M}(\text{source}), \text{target}) \tag{8}$$

where $\mathbb{M}(\cdot)$ denotes the sentence simiplificaton model, source is the input sentencem and target is the ground-truth simplified sentence.

$$\ell_{\text{idiom}} = \text{CrossEntropy}(\mathbb{M}(\text{idiom}), \text{explanation}) \tag{9}$$

where idiom is the idiomatic expression encountered in the source sentence, and explanation is the corresponding simplified version or explanation of the idiom.

$$\ell_{\text{IAS}} = \ell_{\text{sentence}} + \ell_{\text{idiom}} \tag{10}$$

### 3.3 RISS Module

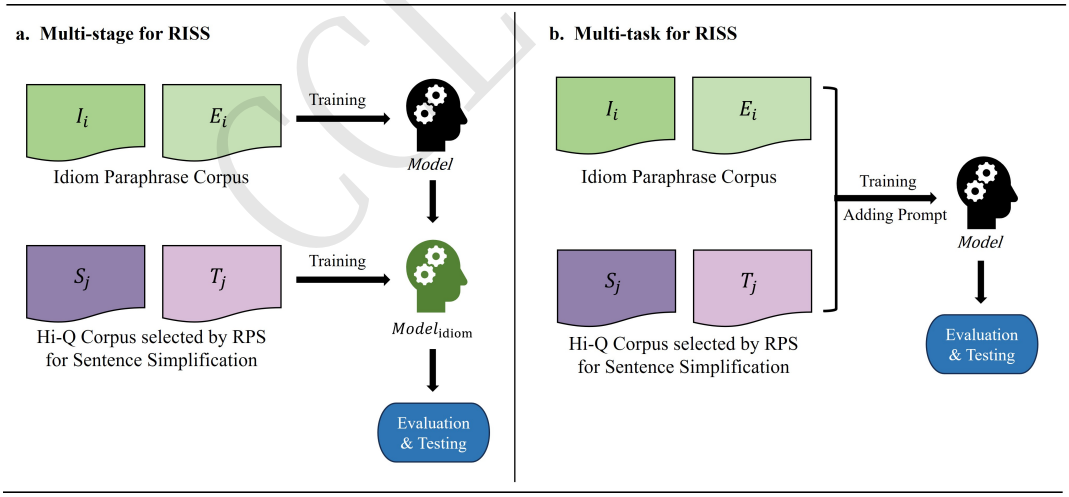



Figure 4: The RISS module architecture with multi-stage (left) and multi-task (right) strategies, where $I_i$, $E_i$ denote idioms and their corresponding explanations respectively, and $S_j$, $T_j$ denote source and target sentences selected by RPS respectively.

The RISS module (Figure 4) integrates RPS and IAS to address the challenges of limited parallel corpora and idiom prevalence in Chinese sentence simplification. RISS employs two training strategies:

**Multi-Stage Training:**

1. Pre-train the model using the IAS module on the idiom simplification dataset.
2. Fine-tune the pre-trained model using the high-quality paraphrase pairs selected by the RPS module.

**Multi-Task Approach with Prompt:**

1. Prepare the training data by adding task-specific prompts to the input sentences. For the idiom simplification task, the input prompt is: "请解释以下成语的意思：" (Please explain the meaning of the following idiom:) For the sentence simplification task, the input prompt is: "请简化以下句子：" (Please simplify the following sentence:)
2. Train the model simultaneously on both tasks using the prompted data. Task-specific prompts help the model distinguish between tasks and apply appropriate simplification strategies.

RISS offers a flexible and effective framework for Chinese sentence simplification by combining RPS and IAS through multi-stage or multi-task training.

## 4 Experiment Design

In this section, we present the experimental setup for evaluating the proposed RISS framework. We describe the datasets used for training and evaluation, the evaluation metrics, and the baseline methods compared against RISS. And we train all the models using implementation of Bart-base-chinese (Lewis et al., 2020).

### 4.1 Dataset

We use the following datasets for training and evaluating our Chinese sentence simplification framework:

**Paraphrase for RPS**. We employ the PKU Paraphrase Bank (Zhang et al., 2019) as the source of paraphrase pairs for the RPS module. This dataset contains 509,832 sentence pairs extracted from 95 translations of 40 novels, providing a rich and diverse set of paraphrase examples.

**Idiom for IAS**. For the IAS module, we utilize the Chinese Idiom Paraphrase (CIP) dataset (Qiang et al., 2022). CIP consists of 105,530 idiomatic expressions along with their paraphrases, divided into training (95,530 pairs), validation (5,000 pairs), and testing (5,000 pairs) sets. We use the training set for our idiom simplification experiments. Note that the CIP already provides the position of the idiom explanation in the target sentence corresponding to the idiom in the source sentence. Consequently, training an additional model to identify idiom explanations is unnecessary. For more details about locating idioms in calculating idiom-specific loss, please refer to Appendix D.

**Evaluation Datasets:** We evaluate the performance of RISS and the baseline methods on two recently released Chinese sentence simplification datasets:

- **CSS** (Sun et al., 2023) is derived from the PFR Chinese corpus and contains 288 manually simplified sentences for few-shot learning and 766 human simplifications (two per original sentence) for evaluation.
- **MCTS** (Chong et al., 2023) is constructed from the Penn Chinese Treebank (CTB) and includes 366 and 357 original sentences for validation and testing, respectively, each with five human simplifications.

Table 1 provides a statistical summary of the CSS and MCTS datasets.

CCL 2024

| Datasets | Train | | Valid | | Test | |
|---|---|---|---|---|---|---|
| | Source | Target | Source | Target | Source | Target |
| CSS | 288 | 288*1 | - | - | 383 | 383*2 |
| MCTS | - | - | 366 | 366*3 | 357 | 357*5 |

Table 1: The statistical analysis of CSS and MCTS. Source denotes original sentences, Target denotes manual simplified sentences and _*$x$ denotes each original sentence has $x$ manually simplified versions.

### 4.2 Evaluation Metrics

To assess the quality of the simplified sentences generated by RISS and the baseline methods, we employ the following automatic evaluation metrics:

**SARI** (Xu et al., 2016) is a commonly used metric for evaluating sentence simplification systems. It calculates the F1 scores for three operations (addition, copying, and deletion) by comparing the system output against the input sentence and multiple human-generated references. Following Wan et al. (2023), we report SARI scores at both character and word levels.

**BERT-Score** (Zhang et al., 2019) measures the semantic similarity between the system output and human references using contextualized embeddings from pre-trained BERT models. We use the bert-base-chinese model from the Hugging Face library to compute BERT-Score.

**BLEU** (Papineni et al., 2002) is a precision-based metric originally designed for machine translation evaluation. It calculates the n-gram overlap between the system output and human references. We use the implementation provided by the NLTK library.

### 4.3 Baselines

We compare the performance of RISS against the following baseline methods:

- **Identity** The original sentences are used as simplification output without any modifications.
- **Truncation** The original sentences are truncated to the first 80% of their words.
- **Direct Back-Translation** (Lu et al., 2021) This method uses back-translation to perform unsupervised sentence simplification, with English as the pivot language.
- **Translate Training** (Chong et al., 2023; Sun et al., 2023) The WikiLarge English sentence simplification dataset (Zhang and Lapata, 2017) is translated into Chinese and used to train a Chinese simplification model.
- **Cross-Lingual Pseudo** (Chong et al., 2023) This method constructs pseudo-parallel data for Chinese sentence simplification by translating a Chinese corpus into English, simplifying the English sentences using an existing model, and then translating the simplified sentences back into Chinese.
- **mT5-large** (Sun et al., 2023)Directly finetune with the additional dataset from CSS.
- **GPT-3.5-turbo** (Feng et al., 2023) A large language model fine-tuned for sentence simplification, applied in a zero-shot or few-shot setting.

## 5 Results and Analysis

In this section, we present the experimental results of the proposed Readability-guided Idiom-aware Sentence Simplification (RISS) framework and compare its performance against the baseline methods on the CSS and MCTS datasets. We also conduct an ablation study to investigate the individual contributions of the RPS and IAS modules.

CCL 2024

### 5.1 Main Result

Table 2 presents the main results of our experiments on the CSS and MCTS datasets. We report the best-performing RISS variant (multi-stage or multi-task) for each dataset and compare it with the baseline methods. The main difference between "Unsupervised (zero-shot)" and "Supervised (few-shot)" is whether the model is fine-tuned using the CSS training set. The training data for both settings include the "idiom paraphrase corpus" and "high-quality corpus selected by RPS" (Figure 4). For the MCTS dataset, which lacks training data, we use the CSS training set for fine-tuning in the supervised experiments.

| Method | CSS | | | | MCTS | | | |
|---|---|---|---|---|---|---|---|---|
| | $Sari_{char}$ | $Sari_{word}$ | *B.S.* | *BLEU* | $Sari_{char}$ | $Sari_{word}$ | *B.S.* | *BLEU* |
| Identity | 29.08 | 27.61 | 0.88 | 88.77 | 25.06 | 22.37 | 0.90 | 84.75 |
| Truncation | 32.95 | 33.18 | 0.79 | 76.36 | 25.21 | 21.85 | 0.87 | 61.6 |
| Gold Reference | 46.72 | 45.71 | 0.96 | 65.31 | 50.12 | 48.11 | 0.90 | 61.62 |
| **Unsupervised (Zero-Shot) Method** | | | | | | | | |
| Lu et al. (2021) | 36.27 | 33.39 | - | 63.47 | - | 40.37 | - | 48.72 |
| Translate Training (2023) | 36.02 | 34.44 | - | 71.41 | - | 28.30 | - | **82.20** |
| Cross-Lingual Pseudo (2023) | - | - | - | - | - | 38.49 | - | 63.06 |
| Gpt-3.5-turbo (2023) | 31.95 | 28.92 | 0.83 | 42.22 | - | 42.39 | - | 49.22 |
| **RISS$_{\text{multi-stage}}$** | 40.95 | 39.36 | 0.81 | 81.92 | 44.37 | 42.71 | **0.90** | 79.94 |
| **RISS$_{\text{multi-task}}$** | **41.68** | **40.52** | **0.87** | **84.05** | **46.23** | **44.36** | **0.90** | 80.83 |
| **Supervised (Few-Shot) Method** | | | | | | | | |
| mT5-large (2023) | 37.57 | 35.97 | - | 74.71 | - | - | - | - |
| Gpt-3.5-turbo (2023) | 39.32 | 36.57 | 0.85 | 60.57 | - | - | - | - |
| **RISS$_{\text{multi-stage+labeled data}}$** | **44** | **43.1** | **0.88** | 88.48 | 47.41 | 45.94 | **0.90** | 79.94 |
| **RISS$_{\text{multi-task+labeled data}}$** | 43 | 42.35 | **0.88** | **88.59** | **48.49** | **46.42** | **0.90** | **81.12** |

Table 2: Main results of the experiment. For a comprehensive analysis, we calculate the missing metrics for publicly available system-generated results provided by previous studies; otherwise, we use "-" to fill in the missing values. **Bold** indicates the best result, and underline indicates the second-best result. For detailed experiment parameters, please see Appendix B.

**RISS Unsupervised Results**. RISS achieves state-of-the-art performance on both datasets in the unsupervised setting. On CSS, RISS obtains a 6.08 SARI point improvement over the previous best unsupervised method and even outperforms the best supervised method by 2.36 points. On MCTS, RISS surpasses the previous best unsupervised method by 1.97 SARI points. Moreover, RISS maintains high semantic similarity with the original sentences, as evidenced by BERT-Scores above 0.8 on both datasets.

The multi-task variant of RISS consistently outperforms the multi-stage variant in the unsupervised setting. We hypothesize that the multi-task approach allows the model to better leverage knowledge from the large-scale idiom paraphrase dataset while simultaneously learning to perform sentence simplification, enhancing its ability to handle idiomatic expressions.

**RISS Supervised Results**. When incorporating labeled simplification data, RISS further advances the state-of-the-art on both datasets. Although the performance of the multi-stage and multi-task variants is not consistent across CSS and MCTS, both variants achieve significant improvements over the baselines, demonstrating the effectiveness of our proposed framework.

### 5.2 Ablation Study

Table 3 presents an ablation study investigating the individual contributions of the RPS and IAS modules.

**RPS:** In both supervised and unsupervised settings, the RPS module improves overn Raw Paraphrase on both datasets, with the highest character-level SARI score improvement reaching 4.09 points. This highlights the importance of mining high-quality paraphrase pairs.

CCL 2024

| Method | CSS | | | | MCTS | | | |
|---|---|---|---|---|---|---|---|---|
| | $Sari_{char}$ | $Sari_{word}$ | *B.S.* | *BLEU* | $Sari_{char}$ | $Sari_{word}$ | *B.S.* | *BLEU* |
| **Unsupervised (Zero-Shot) Method** | | | | | | | | |
| Raw Paraphrase | 36.45 | 35.22 | 0.806 | 86.68 | 41.27 | 39.8 | **0.908** | 81.1 |
| RPS | 40.54 | 38.98 | 0.867 | 82.51 | 43.89 | 42.29 | 0.90 | 78.24 |
| $IAS_{\text{without-loss}}$ | 36.23 | 35.39 | 0.874 | 89.17 | 45.84 | 43.72 | 0.90 | **81.73** |
| $IAS_{\text{with-loss}}$ | 36.56 | 35.89 | **0.878** | **89.66** | **46.31** | **44.76** | 0.901 | 81.18 |
| $RISS_{\text{multi-stage}}$ | 40.95 | 39.36 | 0.811 | 81.92 | 44.37 | 42.71 | 0.899 | 79.94 |
| $RISS_{\text{multi-task}}$ | **41.68** | **40.52** | 0.873 | 84.05 | 46.23 | 44.36 | 0.899 | 80.83 |
| **Supervised (Few-Shot) Method** | | | | | | | | |
| Bart | 40.21 | 39.39 | 0.886 | 88.13 | 44.8 | 43.43 | 0.899 | 81.48 |
| Raw Paraphrase | 41.32 | 40.33 | 0.868 | **89.45** | 45.13 | 44.28 | **0.905** | 81.49 |
| RPS | 43.76 | 42.68 | 0.87 | 87.84 | 46.47 | 45.48 | 0.903 | 80.94 |
| $IAS_{\text{without-loss}}$ | 41.51 | 40.73 | 0.877 | 88.34 | 47.37 | 45.71 | 0.903 | **81.68** |
| $IAS_{\text{with-loss}}$ | 42.99 | 42.26 | **0.902** | 88.63 | 47.62 | 46.11 | 0.895 | 81.15 |
| $RISS_{\text{multi-stage}}$ | **44** | **43.1** | 0.875 | 88.48 | 47.42 | 45.94 | 0.903 | 81.57 |
| $RISS_{\text{multi-task}}$ | 43 | 42.35 | 0.876 | 88.59 | **48.49** | **46.42** | 0.904 | 81.12 |

Table 3: Ablation Study on CSS and MCTS. **Bold** indicates the best result, and underline indicates the second-best result.

**IAS:** The IAS module also consistently improves SARI scores, although the gains are smaller compared to RPS. Interestingly, even without the RPS module or labeled simplification data, IAS achieves the best unsupervised SARI score on MCTS, outperforming some supervised methods. We attribute this to the higher proportion of idioms in MCTS (14.3%) compared to CSS (10.4%), suggesting that idiom paraphrasing is particularly effective for datasets with a higher prevalence of idiomatic expressions.

**RISS** The full RISS framework, combining RPS and IAS through either multi-stage or multi-task training, achieves the best or second-best SARI scores in most settings, confirming the effectiveness of our proposed approach.

Previous study has pointed out that the BLEU metric might be misleading in assessing sentence simplification (Sulem et al., 2018), and its shortcomings are particularly evident in Chinese (Sun et al., 2023; Chong et al., 2023). Furthermore, BERTScore primarily measures semantic preservation and dose not effectively indicate the extent of simplification. When evaluating simplification systems, focus should be on SARI scores. We provide additional evaluation metrics to offer more comprehensive analysis. Although IAS achieves higher BERTScore and BLEU scores in the ablation study, the design of the RPS module remains essential for improving overall simplification performance, as evidenced by the SARI scores.

## 6 Conclusion

In this work, we propose Readability-guided Idiom-aware Sentence Simplification (RISS), a novel framework that addresses two major challenges in Chinese sentence simplification: the scarcity of labeled parallel corpora and the prevalence of idioms. RISS combines two key components: Readability-guided Paraphrase Selection (RPS), which mines high-quality paraphrase pairs from a large corpus based on readability differences and linguistic similarities, and Idiom-aware Simplification (IAS), an idiom paraphrasing model that leverages an idiom-specific loss function to enhance the comprehension and simplification of idiomatic expressions. By integrating RPS and IAS through multi-stage and multi-task learning strategies, RISS achieves state-of-the-art performance on two Chinese sentence simplification datasets, even without using labeled simplification data. Furthermore, when fine-tuned on a small labeled dataset, RISS demonstrates further improvements, advancing the state-of-the-art in Chinese sentence simplification. We hope that our work will inspire future research in low-resource sentence simplification and contribute to making written content more accessible to diverse audiences.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1183-1200, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

## 7 Limitation

While RISS demonstrates state-of-the-art performance in Chinese sentence simplification, there are opportunities for further improvement. The quality scoring algorithm in the RPS module, although effective, could potentially be enhanced to identify and retain an even larger proportion of high-quality paraphrase pairs. Additionally, due to time constraints, we relied solely on automatic evaluation metrics to assess the quality of the system-generated sentences. While these metrics provide valuable insights, a human evaluation would be necessary to comprehensively evaluate the readability, fluency, and meaning preservation of the simplified sentences produced by RISS.

## 8 Acknowledgement

This work is partially supported by Guangzhou Science and Technology Plan Project (202201010729), Major Program of National Social Science Foundation of China (18ZDA295), and Guangdong Social Science Foundation Project (GD24CWY11). We thank the anonymous reviewers for their helpful comments and suggestions.

## References

Emil Abrahamsson, Timothy Forni, Maria Skeppstedt and Maria Kvist. 2014. Medical text simplification using synonym replacement: Adapting assessment of word difficulty to a compounding language. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pages 57-65.

Fernando Alva-Manchego, Louis Martin, Carolina Scarton and Lucia Specia. 2019. EASSE: Easier Automatic Sentence Simplification Evaluation. Conference on Empirical Methods in Natural Language Processing.

Suha S. Al-Thanyyan and Aqil M. Azmi. 2021. Automated text simplification: A survey. *ACM Computing Surveys*, 54(2).

Elron Bandel, Ranit Aharonov, Michal Shmueli-Scheuer, Ilya Shnayderman, Noam Slonim and Liat Ein-Dor. 2022. Quality Controlled Paraphrase Generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

Steven Bird. 2006. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 69–72.

Samuel A. Bobrow and Susan M. Bell. 1973. On catching on to idiomatic expressions. *Memory and cognition*.

Alison Chi, Li-Kuang Chen, Yi-Chen Chang, Shu-Hui Lee and Jason S. Chang. 2023. Learning to paraphrase sentences to different complexity levels.

Ruining Chong, Luming Lu, Liner Yang, Jinran Nie, Shuhan Zhou, Yaoxin Li and Erhong Yang. 2023. MCTS: A Multi-Reference Chinese Text Simplification Dataset. ArXiv, abs/2306.02796.

Liam Cripwell, Joël Legrand and Claire Gardent. 2023. Simplicity Level Estimate (SLE): A Learned Reference-Less Metric for Sentence Simplification. In *Conference on Empirical Methods in Natural Language Processing*.

Yue Dong, Zichao Li, Mehdi Rezagholizadeh and Jackie Chi Kit Cheung. 2019. EditNTS: An neural programmer-interpreter model for sentence simplification through explicit editing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3393-3402.

Yutao Feng, Jipeng Qiang, Yun Li, Yunhao Yuan, and Yi Zhu. 2023. Sentence simplification via large language models. arXiv preprint arXiv:2302.11957.

Wenjie Hao, Hongfei Xu, Deyi Xiong, Hongying Zan, and Lingling Mu. 2022. ParaZh-22M: a Large-Scale Chinese Parabank via Machine Translation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3885–3897.

CCL 2024

Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. 2020. Neural CRF model for sentence alignment in text simplification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7943-7960.

Reno Kriz, Joao Sedoc, Marianna Apidianaki, Carolina Zheng, Gaurav Kumar, Eleni Miltsakaki, and Chris Callison-Burch. 2020. Complexity-weighted loss and diverse reranking for sentence simplification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3137-3147.

Levenshtein, V. I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady* Vol. 10, No. 8, pages 707-710.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Shuang Li, Jiangjie Chen, Siyu Yuan, Xinyi Wu, Hao Yang, Shimin Tao, and Yanghua Xiao. 2023. Translate Meanings, Not Just Words: IdiomKB's Role in Optimizing Idiomatic Translation with Language Models. ArXiv, abs/2308.13961.

Ho Hung Lim, Tianyuan Cai, John S. Y. Lee, and Meichun Liu. 2022. Robustness of Hybrid Models in Cross-domain Readability Assessment. In *Australasian Language Technology Association Workshop*.

Dawei Lü, Xinying Qiu, and Yi Cai. 2020. Sentence-Level Readability Assessment for L2 Chinese Learning. In *Lecture Notes in Computer Science*,pages 381–92.

Xinyu Lu, Jipeng Qiang, Yun Li, Yunhao Yuan, and Yi Zhu. 2021. An unsupervised method for building sentence simplification corpora in multiple languages. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 227–237.

Yuan Ma, Sandaru Seneviratne, and Elena Daskalaki. 2022. Improving Text Simplification with Factuality Error Detection. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability*.

Mounica Maddela, and Wei Xu. 2018. A word-complexity lexicon and a neural readability ranking model for lexical simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3749-3760.

Mounica Maddela, Fernando Alva-Manchego, and Wei Xu. 2021. Controllable text simplification with explicit paraphrasing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3536-3553.

Louis Martin, Samuel Humeau, Pierre-Emmanuel Mazaré, Éric de La Clergerie, Antoine Bordes, and Benoît Sagot. 2018. Reference-less quality estimation of text simplification systems. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, pages 29–38, Tilburg, the Netherlands. Association for Computational Linguistics.

Louis Martin, Angela Fan, Eric de la Clergerie, Antoine Bordes, and Beno ıt Sagot 2022. MUSS: Multilingual Unsupervised Sentence Simplification by Mining Paraphrases. In *International Conference on Language Resources and Evaluation*.

Matej Martinc, Senja Pollak, and Marko Robnik–Šikonja. 2021. Supervised and Unsupervised Neural Approaches to Text Readability. In *Computational Linguistics*, pages 141–179.

Shardlow, M. 2014. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications*, 4(1), 58-70.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Sarah E Petersen and Mari Ostendorf. 2007. Text simplification for language learners: a corpus analysis. In *Workshop on speech and language technology in education*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems, 32.

CCL 2024

Paetzold, G., and Specia, L. 2017. Lexical simplification with neural ranking. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 34-40.

Jipeng Qiang, Yang Li, Chaowei Zhang, Yun Li, Yunhao Yuan, Yi Zhu, and Xindong Wu. 2022. Chinese Idiom Paraphrasing. In *Transactions of the Association for Computational Linguistics*, volume 11, pages 740-754.

Jipeng Qiang, Kang Liu, Qiang, Yun Li, Yunhao Yuan, Yi Zhu and Kaixun Hua. 2023. Multilingual Lexical Simplification via Paraphrase Generation. arXiv preprint arXiv:2307.15286.

Luz Rello, Ricardo Baeza-Yates and Stefan Bott Horacio Saggion. 2013. Simplify or help? Text simplification strategies for people with dyslexia. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*, pages 1-10.

Michael J. Ryan, Tarek Naous, and Wei Xu. 2023. Revisiting non-English Text Simplification: A Unified Multilingual Benchmark. ArXiv, abs/2305.15678.

Victor SANH, Lysandre DEBUT, Julien CHAUMOND, and Thomas WOLF. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. ArXiv, abs/1910.01108.

Sulem, E., Abend, O., and Rappoport, A. 2018. Semantic structural evaluation for text simplification. In *Proceedings of the NAACL-HLT 2018*, pages 685–696.

Renliang Sun, Wei Xu and Xiaojun Wan. 2023. Teaching the pre-trained model to generate simple texts for text simplification. arXiv preprint arXiv:2305.12463.

Renliang Sun, Zhixian Yang and Xiaojun Wan. 2023. Exploiting Summarization Data to Help Text Simplification. ArXiv, abs/2302.07124.

Renliang Sun, Shiping Yang and Xiaojun Wan. 2023. A New Dataset and Empirical Study for Sentence Simplification in Chinese. ArXiv, abs/2306.04188.

Sanja Stajner. 2021. Automatic text simplification for social good: Progress and challenges. In *Findings of the Association for Computational Linguistics: ACL IJCNLP 2021*, pages 2637–2652.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Wei Xu, Chris Callison-Burch and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3, 283-297.

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo H. Paetzold, Lucia Specia, Sanja Štajner, Anaıs Tack and Marcos Zampieri. 2018. A report on the complex word identification shared task 2018. In *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 66-78.

Bowei Zhang, Weiwei Sun, Xiaojun Wan, and Zongming Guo. 2019. PKU Paraphrase Bank: A Sentence-Level Paraphrase Corpus for Chinese. In *Lecture Notes in Computer Science*, pages 814–826.

Kaizhong Zhang and Dennis Shasha. 1989. Simple fast algorithms for the editing distance between trees and related problems. In *SIAM journal on computing*, 18(6):1245–1262.

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. arXiv preprint arXiv:1703.10931.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with BERT. CoRR, abs/1904.09675.

Yanbin Zhao, Lu Chen, Zhi Chen, and Kai Yu. 2020. Semi-supervised text simplification with back-translation and asymmetric denoising autoencoders. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05), pages 9668-9675.

CCL·2024

## Appendix A Glossary of Abbreviations

| | |
|---|---|
| RISS | Readibility-guided Idiom-aware Sentence Simplification |
| RPS | Readabilituy-guided Paraphrase Selection |
| IAS | Idiom-aware Simplification |
| RISS_multi-stage | A multi-stage training strategy where the model is first pre-trained using the idiom paraphrase dataset (IAS module) and then trained on the mined paraphrase dataset (RPS module). |
| RISS_multi-task | A multi-task training strategy where the idiom paraphrase dataset and the mined paraphrase dataset are used together for training. Each sentence is prefixed with a corresponding prompt. |
| RISS_multi-stage+labeled data | A multi-stage training strategy similar to RISS_multi-stage, but with an additional fine-tuning step using the standard labeled dataset after saving the best model from the pre-training stage. |
| RISS_multi-task+labeled data | The best model from the multi-task learning strategy (RISS_multi-task) is saved and then fine-tuned using the standard labeled dataset. |
| Raw Paraphrase | The publicly available large-scale paraphrase corpus is directly used for pre-training without any quality filtering. |
| RPS (as in Table 3) | The publicly available large-scale paraphrase corpus is quality-filtered using the proposed RPS module, and the remaining high-quality paraphrase data is used for pre-training. |
| IAS_without-loss | The model is pre-trained using the publicly available idiom paraphrase corpus without introducing the idiom-specific loss. |
| IAS_with-loss | The model is pre-trained using the publicly available idiom paraphrase corpus while introducing the idiom-specific loss to further enhance the model's ability to understand and handle idioms. |

## Appendix B Implementation Details

In our work, we train all models using PyTorch and the implementation of Bart-base-chinese from Hugging Face. All models are trained on an NVIDIA GeForce RTX 3080, using the Trainer provided by the Transformers library as our experimental environment. Table 4 presents our specific experimental hyper-parameters. Following Sun et al.(2023), we filtered out sentences with less than 30 tokens when processing the paraphrase corpus. Additionally, we removed sentence pairs with an RSRS difference lower than 0.1, an empirical value, to prevent inadequate simplification.

CCL 2024

| Hyper-Parameter | Value | | |
|---|---|---|---|
| | Multi-stage Pretrain | Multi-task Pretrain | SS Finetune |
| Max source length | 256 | 256 | 256 |
| Max target length | 256 | 256 | 256 |
| Batch size | 4 | 4 | 4 |
| Learning rate | 2e-5 | 2e-5 | 2e-5 |
| Grad. accu. steps | 2 | 2 | 2 |
| Evaluation strategy | steps | steps | steps |
| Eval step | 500 | 1000 | - |
| Optim metric | Sari | Sari | Sari |
| Max steps/epoch | 50000 | 100000 | 25 |

Table 4: Detailed hyper-parameter settings for multi-stage (i.e., using idiom paraphrase or the PKU paraphrase for pretraining respectively.), multi-task pretrain, and the few-shot finetune in our work. For the other hyper-parameters, we used the default values provided by the Transformers library.

## Appendix C Statistical Analysis of idioms in CSS and MCTS

| Dataset | Train | Val | Test |
|---|---|---|---|
| CSS | 0.347% | 10.44% | - |
| MCTS | - | 14.48% | 14.28% |

Table 5: Statistical analysis of idioms in CSS and MCTS, where ”-” indicates that the corresponding data split is not available in the dataset, making it impossible to calculate the idiom statistics.

## Appendix D Locating idiom position for calcualting idiom-specific loss

| | |
|---|---|
| Source | 如果我们无法找到内心的平和，那么去别处寻找无疑是缘木求鱼。[SEP] 如果我们无法找到内心的平和，那么去别处寻找无疑是 <extra_id_0>。 |
| Target | <extra_id_0> 没有意义的 [null]<extra_id_1> |

Table 6: A specific example from CIP, where ”<extra_id_0>” in Source denotes the idiom ” 缘木求鱼” in the original sentence, and followed by ”<extra_id_0>” in Target is the corresponding explanation ” 没有意义的”. ”[null]” in this context does not carry any particular meaning.

In the source sentence, <extra_id_0> represents the idiom ” 缘木求鱼”, while in the target sentence, the explanation ” 没有意义的” immediately follows <extra_id_0>.

When calcuating the idiom-specific loss, we do not need to train an additional model to identify idiom explanations. Instead, we can obtain the expected idiom explanation and its corresponding index by writing simple splitting, matching, and replacement code. We also construct a dictionary with idioms as keys and the corresponding explanation indices as values for the additional idiom loss during training.

The detailed process is as follows (using the example in Table 6):

CCL 2024

1. Obtain the target sentence and the corresponding idiom explanation:
Extract the sentence containing the idiom mask from the source using [SEP] as the separator, and extract the corresponding explanation from the target. Replace "<extra_id_0>" with the explanation to obtain the target sentence with the idiom explanation.

2. Obtain the indices of the idiom explanation in the target sentence:
Encode the target sentence using a tokenizer to get a series of indices in the PLM vocabulary. Match the indices of the idiom explanation in the target sentence to determine its position.

3. During training, precisely locate the idiom explanation position using the dictionary and calculate the cross-entropy loss.

In summary, our approach leverages the provided idiom paraphrase dataset to accurately identify and simplify idioms without the need for additional training. The idiom-specific loss is calculated using the indices of the idiom explanation in the target sentence, ensuring the model's ability to handle idioms effectively.

## Appendix E Unsupervised RSRS Metric

The performance of *RSRS* depends on the evaluation model. To determine the best model, we downloaded pre-trained language models (PLMs) with different architectures from Hugging Face[2] and performed the complexity binary prediction problem according to the *RSRS* values on CSS dataset. Based on the prediction results, we chose distilbert-base-multilingual-cased as our *RSRS* discriminant model. Table 7 presents detailed results.

Additionally, we calculated the mean RSRS value for source and reference sentences in CSS dataset, which were 20.00 and 17.99, respectively. A pairwise t-test revealed a significant difference in RSRS between the original and reference sentences ($p < 0.05$), further validating the effectiveness of RSRS.

| PLMs | Accuracy | F1-score |
|---|---|---|
| bert-base-multilingual-uncased | 0.75 | 0.857 |
| bert-base-chinese | 0.51 | 0.676 |
| chinese-roberta-wwm-ext | 0.573 | 0.728 |
| xlm-roberta-base | 0.66 | 0.795 |
| xlm-roberta-base-language-detection | 0.549 | 0.709 |
| **distilbert-base-multilingual-cased** | **0.785** | **0.879** |

Table 7: We evaluated six different PLMs for RSRS on the CSS dataset. The experimental results show that distilbert-base-multilingual-cased achieved an F1-score of 0.879 and an accuracy of 0.785, outperforming the other five models.

[2] https://huggingface.co/

CCL 2024

## Appendix F Statistical Analysis of Our RPS Data and CSS

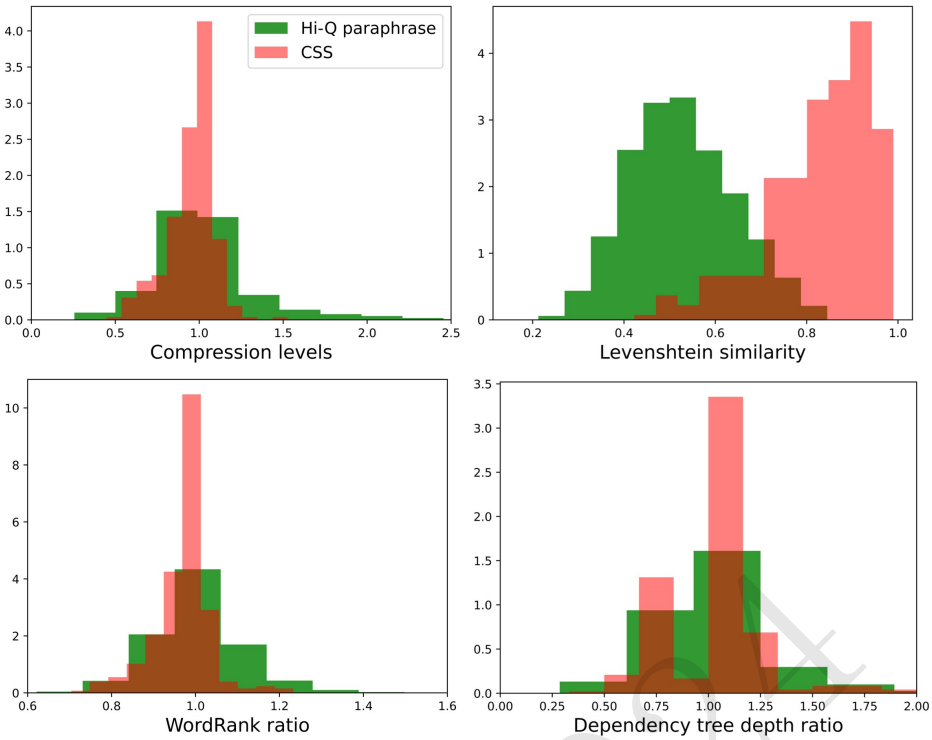



Figure 5: Sentence feature density in the CSS dataset and the data mined by our RPS module. The WordRank ratio is a measure of lexical complexity reduction.

## Appendix G SARI Evaluation Details

**Algorithm 1:** Calculation Process for SARI metric

**Input :** Original sentences List $orig\_sents$, System generated sentences List $sys\_sents$, Number of Reference version $x$, Reference sentences List(s) $refs\_sents_i$, $1 \leq i \leq x$.

**Output:** Sari score on CSS $Sari_{css}$, Sari score on MCTS $Sari_{mcts}$.

$Sari\_list \leftarrow []$;
**for** *idx, orig in enumerate(orig_sents)* **do**
  $orig, sys \leftarrow [orig], [sys\_sents[idx]]$;
  $refs \leftarrow [[refs\_sents_1[idx]], [refs\_sents_2[idx]], ..., [refs\_sents_x[idx]]]$;
  $sari \leftarrow$ compute sari $(orig, sys, refs)$;
  $Sari\_list$.append$(sari)$;
**end**
$Sari_{css} \leftarrow mean(Sari\_list)$;
$Sari_{mcts} \leftarrow$ compute sari $(orig\_sents, sys\_sents, [refs\_sents_1, refs\_sents_2, ..., refs\_sents_x])$;
return $Sari_{css}$, $Sari_{mcts}$

Figure 6: Pseudo code for computing sari metric on CSS and MCTS.

Note that document- and sentence-level inputs can lead to different results even if the model generates the same results. Please refer to the pseudo code in Figure 6.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1183-1200, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

## Appendix H Case Study

| | |
|---|---|
| Original | 朱老太太原本是上海人，有着上海人的优越感，做事井井有条，说话轻声细语。<br>Mrs. Zhu was originally from Shanghai and had the superiority of Shanghai natives, doing things in a well-organized manner, and speaking with a gentle voice. |
| Gpt-3.5-turbo | 朱老太太原本是上海人，做事有条不紊，说话轻声细语，有上海人的优越感。<br>Mrs.Zhu was originally from Shanghai, methodically in her work and spoke with a gentle voice, having the superiority of Shanghai natives. |
| $RISS_{multi-stage}$ | 朱老太太原是上海人，有着上海的优越感，做事很有条理，说话轻声细语。<br>Mrs.Zhu was originally from Shanghai and had the superiority of Shanghai natives, doing things very methodically, and speaking with a gentle voice. |
| $RISS_{multi-task}$ | 朱老太太原本是上海人，有着上海的优越感，做事很有条理，说话很温和。<br>Mrs.Zhu was originally from Shanghai and had the superiority of Shanghai natives, doing things very methodically, and speaking gently. |
| $RISS_{multi-stage+labeled\ data}$ | 朱老太太原本是上海人，有着上海一般人的优越感，做事十分有条理，说话轻声细语。<br>Mrs.Zhu was originally from Shanghai and had the superiority of Shanghai natives, doing things very methodically, and speaking with a gentle voice. |
| $RISS_{multi-task+labeled\ data}$ | 朱老太太原本是上海人。她有着上海的优越感，做事很有条理，说话很轻声。<br>Mrs.Zhu was originally from Shanghai. She has the superiority of Shanghai natives, doing things very methodically, and speaking gently. |
| $Reference_1$ | 朱老太太原本是上海人，做事很有条理，说话轻声细语。<br>Mrs. Zhu was originally from Shanghai, doing things very methodically, and speaking with a gentle voice. |
| $Reference_2$ | 朱老太太是上海人，有着上海人的优越感。她做事井井有条，说话轻声细语。<br>Mrs. Zhu was from Shanghai and had the superiority of Shanghai natives. She did things in a well-organized manner, and speaking with a gentle voice. |

Table 8: Generated example from the CSS, where the red font denotes the part of the original sentence that has been simplified by system.

As shown in Table 8, both unsupervised and supervised RISS can identify and simplify the idiom ” 井井有条” (in a well-organized manner) to ” 有条理” (methodically). Furthermore, the multitask RISS can rephrase ” 轻声细语” (in a gentle voice) as ” 温和” (gently). After incorporating the labeled dataset, RISS further demonstrates clause awareness in its simplifications.

CCL 2024

# A Tone-based Hierarchical Structure of Chinese Prosody

**Ya Li**
Zhejiang Ocean University/ No. 1 Haida Road, Zhoushan, Zhejiang 316022
liyawy@zjou.edu.cn

**Abstract**

In Chinese speech engineering, many projects use a conventional, syllable-based prosodic hierarchy as an underlying framework to process natural or synthesized speech. However, Chinese as a tone language has its own way of expressing prosody, that is, through tonal interaction, especially tone sandhi. By utilizing the capacity of tone as a dual unit of pitch and timing, the present study proposes a tone-based, three-layer-four-level structure for Chinese prosody. The three layers are tone, tone prosody, and intonation, respectively composed of one level of pitch units, two levels of tone prosody units (basic and derived), and one level of intonation units. These four levels of units are used to replace syllable, prosodic word, phonological phrase, and intonational phrase in a conventional hierarchy. Tone prosody units are established based on sizes or types of tone sandhi domains, so when applied to the same clause uttered in Mandarin and Shanghai Wu Chinese, they are timed differently and branched toward different directions at different levels, hence capable of capturing rhythmic and melodic patterns of the two distinctive types of Chinese. Overall, given its theory-friendly design, the proposed structure may be used as a unifying framework in Chinese speech engineering.

## 1 Introduction

In speech engineering, it is notoriously difficult to improve the naturalness of synthesized speech due to multiple linguistic and non-linguistic factors involved. It is especially true for Chinese speech synthesis, because it is hard to find a unifying prosodic framework optimized for the whole Chinese language family, both standard Mandarin Chinese and other Chinese dialects. At present, most models assume a syllable-based prosodic hierarchy developed for non-tonal languages. Such a hierarchy is conventionally divided into eight layers from bottom to top: mora, syllable, foot, prosodic word, clitic, phonological phrase, intonational phrase, and utterance (Nespor and Vogel, 1986). This hierarchy is built around stress foot, hence readily applicable to stress languages like English. For Chinese, according to Cao (2011), a simplified version of this conventional hierarchy is generally assumed, and it has three basic levels, prosodic word (PW), prosodic phrase (PP), and intonational phrase (IP). However, the applicability of such a simplified hierarchy to Chinese is still questionable, because Chinese has tone rather than stress on its syllables and the status of tone is not reflected in this hierarchy.

©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1201-1211, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Alternatively, Zhang (2017) proposed the prosody-syntax sensitivity model of prosodic hierarchy based on tone sandhi domains. He used examples from Xiamen Min Chinese to show that syntactic information directly participates in the whole process of tone sandhi, from domain construction to sandhi execution. His model acknowledges the role of tone sandhi domains in mediating between syntax and prosody, yet it is still based on the conventional, syllable-based hierarchy. The present study, on the other hand, holds that any Chinese prosodic hierarchy should involve tone directly, as it is the most important feature of Chinese speech. It thus proposes a tone-based hierarchical structure for Chinese prosody.

The present paper first introduces research background (Section 2), followed by an illustration of the proposed structure (Section 3). Then, examples of its application are presented (Section 4) and comparisons are made between the proposed structure and other structures (Section 5). Last, implications of the proposed structure for prosody modeling are discussed (Section 6), followed by a brief conclusion (Section 7).

## 2 Research background

Similar to Zhang (2017), who claimed that the Chinese prosodic hierarchy must refer to tone sandhi domains, the present study holds that a tone-based prosodic hierarchy must be built around tone sandhi domains. According to Yan (2010), tone sandhi domains are hierarchically ordered and tone sandhi rules take prosodic units as their domain of application: Some take prosodic word while others take phonological phrase. Nonetheless, this is not always the case with all Chinese dialects.

Worse than all, the notions of word and phrase in Chinese are problematic, as they are very different from those in English: a Chinese *zi*, a mono-syllabic meaning unit, is usually a word by itself. A multi-*zi* sequence, when grammatical markers are lacking, can be used as a word, a phrase, or a clause, depending on the pragmatic context. Zhuang (2015) suggested that a Chinese word should be viewed as a unit across multiple prosodic layers since there is no clear boundary between a word and a phrase. Shi (2017) also proposed a dual unit called Chinese *zi*-word to encompass both *zi* and word. Establishing a tone-based prosodic hierarchy for Chinese may help differentiate *zi*, word, and phrase if each of them has a distinctive tone sandhi pattern. Better yet, Li (2023) in her study of tonal timing proposed a dual pitch-timing structure for tone, which allows tone to serve as a basic prosodic unit instead of syllable. Moreover, Xu (2018) studied English stress and Chinese tone as well as their respective relationship with rhythm and concluded that Chinese rhythm should be tone- rather than syllable-based.

Here there are two notions, melody and timing, to be clarified. According to Hirst (2013), melody refers to a global pitch fluctuation pattern shaped by tone, stress, and intonation and measured by scope and speed of change in pitch height. A faster and larger pitch fluctuation leads to a more melodious speech. Timing is associated with rhythm. Rhythm refers to a recurring timing pattern of prosodic units, a regular duration change caused by syllable type, tone, stress, pause, and speech rate and measured by scope of change in segmental duration. It generally falls into three classes: stress-timing, syllable-timing, and mora-timing (Ladefoged and Johnson, 2011). In theory, they respectively have stress foot, syllable, and mora equally timed during speech. In reality, their timing distinctions are not fully supported by experimental data. Nonetheless, different languages do show different tendencies toward one class or another.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1201-1211, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

For example, Hirst (2013) found that Mandarin Chinese is more syllable-timed and also more melodious than French, and French in turn more syllable-timed and also more melodious than English. In addition, English declarative sentences show an overall falling melody whereas French ones an overall rising melody. Li (2015) conducted a further study of four Chinese dialects, Northern (including Mandarin), Cantonese, Wu, and Min. Her results show that the four dialects all tend to be syllable-timed and melodious, but to different degrees: Cantonese and Northern dialects are more syllable-timed and also more melodious than Wu and Min. She attributed these differences to their distinctive tone inventories and tonal behaviours.

Last, many experimental studies find that different phonetic cues are used to signal boundaries of different prosodic units (Yin, 2020). In Mandarin, an intonational phrase is cued by pitch height decline towards the end, followed by a long pause; and a phonological phrase by final lengthening. Prosodic words, on the other hand, do not have clear boundary cues, but they can be cued by tone sandhi.

## 3 The proposed prosodic structure

The underlying assumption of the proposed structure is that the pitch-timing duality of tone is capable of directly or indirectly reflecting Chinese melodic and rhythmic patterns. In other words, tone alone is sufficient to characterize Chinese prosody. Therefore, the syllable layer in the conventional prosodic hierarchy can be replaced by the tone layer. Also, the syllable-based stress foot can be replaced by a tone-based unit, called a tone prosody unit (TPU) in the present study. A tone sandhi domain is a typical TPU. Since tone sandhi is closely related to morpho-syntactic conditions, a sandhi-domain-based TPU helps mitigate the need to distinguish words from phrases. TPUs can be divided into two types and placed on two levels. There is good reason to differentiate them in terms of levels, which will be explained soon. Now the tone-based prosodic structure can be illustrated below:

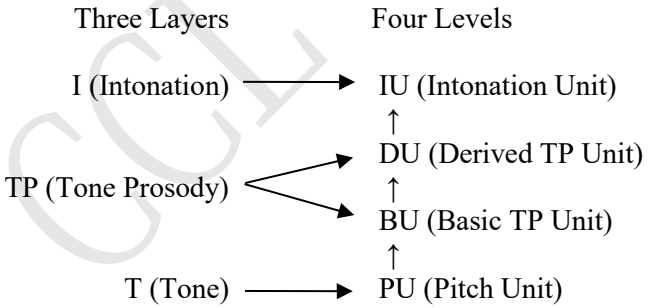



Figure 1. The Tone-based Prosodic Structure of Chinese.

In Figure 1, the tone-based structure has three core layers: T, TP, and I. They are composed of different prosodic units. PU refers to different pitch height components, usually represented by pitch levels, high, mid, and low (H/M/L), or pitch values from 5 to 1 on a five-level scale. For example, the Mandarin fourth tone (T4), a falling tone, is usually represented as HL or 51, indicating that it contains two PUs, H and L. Note that if a BU contains two tones, a tonal interaction process (including tone sandhi) will be evoked to reduce the PU number due to the timing pressure, but the PU in the head position will be retained due to the need to create prosodic prominence (Li, 2023). Take the Mandarin third tone (T3) sandhi for example. If a BU contains a T3.T3 sequence (LLH.LLH), it will undergo the T3 sandhi, turning T3.T3 into a sequence similar

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1201-1211, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

to T2.T3 (LH.LLH). In connected speech, the resulting pitch sequence LH.LLH is often reduced to L**H**L, where the central **H** in the head position is retained.[2]

Next, the structure has BU and DU in the same TP layer but on different levels. They are in the same layer because they both can be domains of tone sandhi and share some structural features; for example, they both can be recursively built by adjoining another tone, a process similar to cliticization. Their division, on the other hand, is for the purpose of accommodating tone sandhi domains of different sizes or types. Later, examples will be given to show their differences. Generally, they encompass the conventionally defined prosodic word, clitic, and phonological phrase, but there is no strict one-on-one relationship between BU and prosodic word as well as between DU and phonological phrase. Last, the upper level IU is equivalent to the conventionally defined intonational phrase. Overall, the proposed structure is aimed to capture Chinese melodic and rhythmic patterns through interactions of different units both within the same or across different layers. Now Mandarin examples are presented to show how the four levels of prosodic units are formed:

| | |
|---|---|
| (a)BU as a tone sandhi domain: | (T3.T3)$_{BU}$ → T2.T3<br>*shui.guo*, 'fruit' |
| (b) BU' (Recursive BU): | ((T3.T3)$_{BU}$T0)$_{BU'}$ → T2.T3.T0<br>*shui.guo.de,* 'of fruit' |
| (c)DU: | [(T3.T3)$_{BU}$(T3.T3)$_{BU}$]$_{DU}$ → T2.T3-T2.T3<br>*shui.guo-pin.zhong,* 'fruit variety' |
| (d) DU' (Recursive DU): | [[(T3.T3)$_{BU}$(T3.T3)$_{BU}$]$_{DU}$T0]$_{DU'}$ → T2.T3-T2.T3.T0<br>*shui.guo-pin.zhong.de,* 'of fruit variety' |
| (e)IU: | {[(T3.T3)$_{BU}$(T3.T3)$_{BU}$]$_{DU}$[(T3.T3)$_{BU}$(T3.T3)$_{BU}$]$_{DU}$}$_{IU}$ → T2.T3-T2.T3-T2.T3-T2.T3<br>*shui.guo-pin.zhong-zhi.you-ji.zhong,* 'There are only a few fruit varieties.' |

Table 1. Formation of Different Levels of Prosodic Units.

In Table 1, the BU in (a) is formed on a Mandarin T3 sandhi domain. The BU in (b) forms a BU' (recursive BU) with T0 (the light tone), a clitic-like tone often carried by a handful of functional words. Two BUs in (c) form a typical DU on the higher level. The DU in (d) forms a DU' (recursive DU) with T0. Note that BU and DU differ only in size. Last, two DUs in (e) form a typical IU on the upper level.

The T3 sandhi applies not only in BUs as shown above but also across adjacent units as shown below:

| | |
|---|---|
| (a) Sandhi applied to a compound structure: | [(T3.T3)$_{BU}$(T3)$_{BU}$]$_{DU}$→[T2.T3-T3]$_{DU}$→T2.T2-T3<br>*yu.san-chang,* 'an umbrella factory' |
| (b)Sandhi applied to a subject-predicate structure: | {[(T3.T3)$_{BU}$]$_{DU}$[(T3)$_{BU}$]$_{DU}$}$_{IU}$→{T2.T3-T3}$_{IU}$→T2.T2-T3<br>*yu.san-xiao,* 'The umbrella is small.' |

Table 2. The Mandarin T3 Sandhi Applying across Adjacent Units.

[2] According to Li (2023), each full tone also has a head position, and the PU in this position indicates the tonal identity. For example, T3 as a dipping tone (L**L**H or 2**1**4) has the head position in the middle and occupied by **L** (bolded), indicating that it is essentially a low tone.

CCL 2024

In Table 2, the phrase in (a) has a compounding structure, so the first two T3s and the last T3 can form two BUs respectively. Next, the two BUs form a DU. The T3 sandhi applies in the first BU and then across the two BUs. The clause in (b) has a subject-predicate structure, so two BUs can further form two DUs respectively. Then the two DUs form an IU. Likewise, the T3 sandhi applies in the first BU and then across the two DUs. The two examples indicate that the T3 sandhi is in essence prosody-based, but its applying order must make reference to morpho-syntactic structures.

Next, Shanghai Wu examples are presented to show how the proposed units are formed based on different types of tone sandhi. According to Li (2017), Shanghai Wu has prosodic words ranging from one to five syllables, and they usually form a tone sandhi domain to which a so-called "wide-application" type of sandhi applies. There is also a contrasting type called "narrow-application", occurring only at the phrase level. Each prosodic word as a wide-application sandhi domain forms a BU and each phrase as a narrow-application domain forms a DU. Therefore, Shanghai Wu tone sandhi is morpho-syntactically rather than prosodically based. The two types of sandhi domains are illustrated below:

| | |
|---|---|
| (a)BU as the wide-application sandhi domain: | $(\underline{34}.\underline{13})_{BU} \rightarrow \underline{33}.\underline{44}$<br>*tshau.ve,* 'stir-fried cooked-rice' |
| (b)DU as the narrow-application sandhi domain: | $[(\underline{34})_{BU}(13)_{BU}]_{DU} \rightarrow \underline{33}$-13<br>*tshau-ve,* 'to stir-fry cooked-rice' |

Table 3. Two Types of Sandhi Domains in Shanghai Wu.

In Table 3, the two-*zi* sequence *tshau ve* [3] carries 34 and 13 originally. In (a), the BU is associated with a prosodic word meaning 'stir-fried cooked-rice', to which the wide-application sandhi applies, turning 34.13 to 33.44. [4] In (b), each tone is associated with a prosodic word and two tones form two BUs respectively. Then the two BUs form a DU, associated with a verbal phrase meaning 'to stir-fry cooked-rice', to which the narrow-application sandhi applies, turning 34-13 to 33-13. The two resulting tone sequences are considered to have the left- and right-heavy prominence respectively (Guo, 2020). The linking relationship between different levels of units in the two Shanghai Wu examples can be further illustrated below:

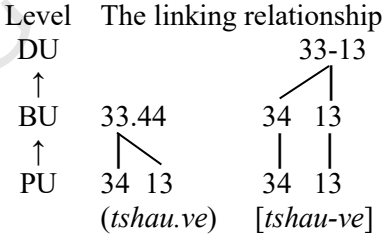



Figure 2. The Linking Relationship between Different Levels of Units in Shanghai Wu Examples.

In Figure 2, the vertical line is linked to a head position and the oblique line to a non-head position. The two-tone BU is linked to the head position on the left and branches to the right, as the wide-application sandhi results in the left-heavy prominence; The two-tone DU is linked to the head position on the right and

[3] Here the Wu spelling (https://www.wugniu.com) is used instead of the Mandarin Pinyin.
[4] 33 and 44 are not citation tones in Shanghai Wu. They only occur as a result of tone sandhi.

branches to the left, as the narrow-application sandhi results in the right-heavy prominence. Therefore, the linking relationship between different levels of units helps reflect prosodic prominence.[5]

## 4 Application of the proposed structure to an utterance

This section shows how the tone-based structure is further applied to a clause, "Once, the North wind and the Sun were disputing which was stronger", uttered by the same female speaker respectively in Mandarin and Shanghai Wu. First, the pitch contour of the Mandarin utterance is displayed below:

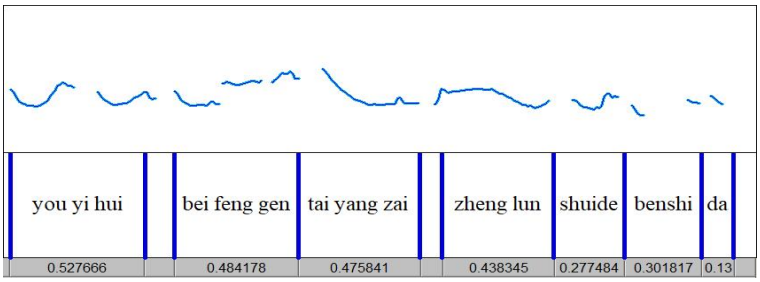



Figure 3. The Pitch Contour of the Mandarin Utterance.

In Figure 3, most units have a size of two or three tones. Also, the first few units are similarly timed, as evidenced by their time intervals around 0.5s. Note that the three-tone unit *you.yi.hui* ('once') has a little longer interval than the following two three-tone units, and the two-tone unit *zheng.lun* ('dispute') has a little shorter interval than the preceding three-tone units. This pattern suggests that both units are on the DU level by themselves, so the final lengthening effect applies to them, making the former longer than and the latter close in length to regular three-tone units.

In terms of melody, the whole contour lowers at the end, as evidenced by a large pitch height difference between the two high-falling T4s, one on *tai* in the word *tai.yang* ('sun') and the other on the final word *da* ('big'). Overall, the contour is relatively melodious. Now the underlying tone-based prosodic structure can be drawn for the Mandarin utterance as follows:

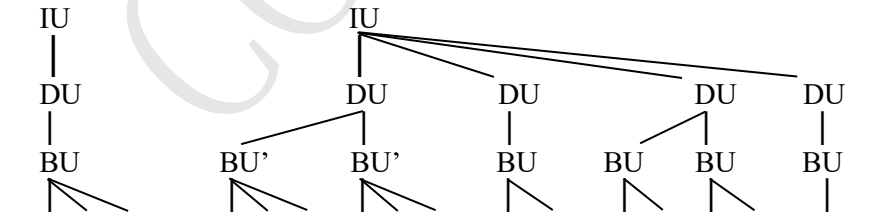



Figure 4. The Tone-based Prosodic Structure for the Mandarin Utterance.

In Figure 4, T1(55), T2(35), T3(214), and T4(51) are full tones and T0 is the light tone. Note that full tones can reduce to T0, a process called neutralization. For example, the high-rising T2 on *yang* in the word

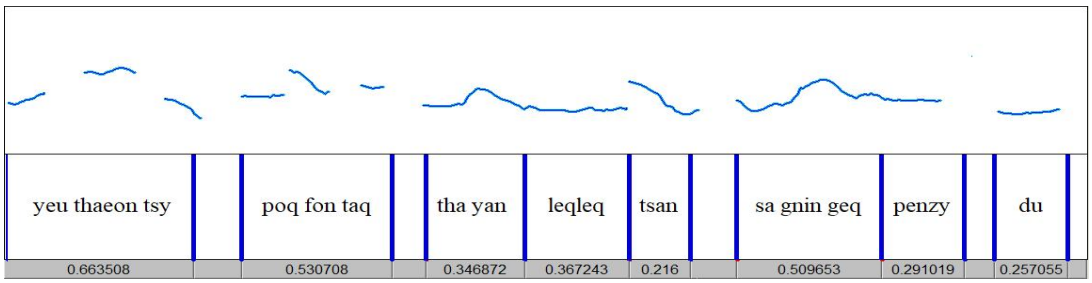
[5] Chen (2000) proposed two similar units, lexical and phrasal MRUs (Minimal Rhythmic Units), to account for different tone sandhi domains. However, he did not discuss their status in a prosodic hierarchy, let alone their linking relationship.

*tai.yang* (‘sun’) is usually neutralized to T0, as evidenced by a large fall instead of rise in the pitch contour shown in Figure 3. The first three tones can be considered as a single BU, because they are on the fixed expression *you.yi.hui* (‘once’). Furthermore, followed by the comma, this BU also forms a DU and an IU successively, as cued by final lengthening plus a relatively long pause (see the gap following *you.yi.hui* in Figure 3). Next, the two tones form a BU respectively in the words, *bei.feng* (‘north wind’), *tai.yang* (‘sun’), *zheng.lun* (‘dispute’), *shui.de* (‘whose’), and *ben.shi* (‘capability’). Then, T1 on *gen* (‘and’) cliticizes to the preceding BU on *bei.feng* (‘north wind’), together forming a BU’; likewise, T4 on the progressive aspect marker *zai* cliticizes to the preceding BU on *tai.yang* (‘sun’), together forming another BU’. Note that BU’s are placed on the same level as BUs for the sake of simplicity.

The two BU’s in the subject position further form a DU, as cued by a short pause (see the gap following *tai.yang.zai* in Figure 3). Next, the two-tone BU on the verb *zheng.lun* (‘dispute’) forms a DU by itself, as cued by final lengthening (see the long tail of its pitch contour in Figure 3). Then the following two BUs form another DU. Note that in the second BU *ben.shi* (T3.T4), the T4 syllable is much longer than the T3 one (in Figure 3, the large gap in the pitch contour belongs to the consonant *sh* in *shi*), indicating that it is lengthened at the phrasal boundary. Finally, the last tone on *da* (‘big’) forms a BU and then a DU successively due to its clause-final position. The last four DUs further form another IU in the utterance.

Also, a prosodic unit, if branched, is right-branching on the BU level, left-branching on the DU level, and right-branching on the IU level. What do these branching directions mean for Mandarin? First, the right-branching BU means that the head position must be on the left, and it cannot accommodate T0. T0 indeed occurs only on the right in a di-syllabic word, a non-head position. Also, if a monosyllabic functional word carries a full tone, the tone cannot be in the head position unless stressed. For example, T4 on the progressive aspect marker *zai* is actually neutralized to T0 just like the preceding T2 on *yang*, as evidenced by a large fall in the pitch contour shown in Figure 3. Therefore, both T0 and an unstressed tone can be treated as a clitic to the preceding tone or BU, rendering the BU or BU’ left-heavy. Next, the left-branching DU reflects the experimental finding that Mandarin phrases often use durational prominence, that is, final lengthening, to cue the right boundary, so they are considered to be right-heavy.[6] Last, the right-branching IU reflects the finding that Mandarin declarative sentences tend to lower the overall pitch level towards the end, hence rendering the IU left-heavy.

Next, the case with the Shanghai Wu utterance is presented, and its pitch contour is displayed below:




[6] Ma (2021) claimed that Mandarin has many compound words carrying a structural stress, whose placement largely depends on the context. The proposed structure places the left- and right-heavy prominence on two levels, suggesting that Mandarin word stress, if present, is more prosodic than lexical in nature, a view along the line with but more explicit than Ma’s (2021).

Figure 5. The Pitch Contour of the Shanghai Wu Utterance.

In Figure 5, most units are sized between two and three tones, but they are not as evenly timed as those in the Mandarin utterance, as evidenced by their varying intervals between 0.2s and 0.7s. Also, the whole contour does not lower much at the end, as evidenced by a small pitch height difference between the two low-rising T6s, one on *yeu* in the first phrase *yeu.thaeon.tsy* (‘once’) and the other on the final word *du* (‘big’). Overall, the Shanghai Wu utterance is less evenly timed and also less melodious than the Mandarin one. Now the underlying tone-based structure can be drawn for the Shanghai Wu utterance as follows:

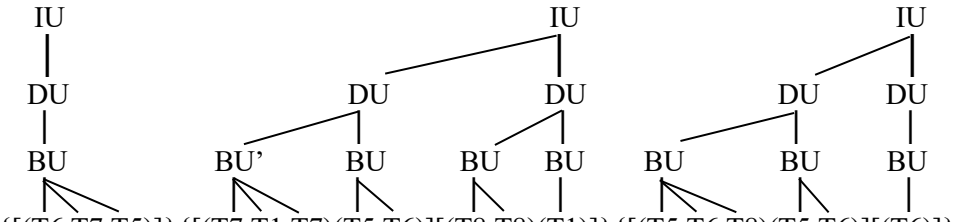



Figure 6. The Tone-based Prosodic Structure for the Shanghai Wu Utterance.

In Figure 6, T1(53), T5(34), T6(13), T7(5, glottalized), and T8(12, glottalized) are all citation tones. The prosodic structure of the Shanghai Wu utterance is clear-cut, since the wide-application sandhi domain is right-branching and left-heavy at the BU level and the narrow-application sandhi domain is left-branching and right-heavy at the DU level. Main structural features of the Shanghai Wu utterance are as follows: 1) following the monosyllabic verb *tsan* (‘dispute’), there is an extra IU boundary occurring with a relatively long pause (see the gap following *tsan* in Figure 5); 2) the progressive aspect marker *leqleq* forms a BU by itself and then a DU with the verb *tsan*; 3) the IU follows the DU in the branching direction and prominence position. They both are right-heavy, which conforms to the previous finding that Wu melody tends to have an overall small pitch fluctuation (Li, 2015).

The table below summarizes the structural and prosodic differences between the two types of Chinese:

| Differences | Mandarin | Shanghai Wu |
|---|---|---|
| Tone sandhi type | T3 Sandhi/T→T0 | Wide-/Narrow-Application |
| Branching direction | BU/DU/IU: Right/Left/Right | BU/DU/IU: Right/Left/Left |
| Pitch fluctuation(melody) | Large(more melodious) | Small(less melodious) |
| TPU length(timing) | More even(more TPU-timed) | Less even(less TPU-timed) |

Table 4. Structural and Prosodic Differences between Mandarin and Shanghai Wu.

In Table 4, the first two rows indicate structural features and the last two prosodic ones. In the last row, the “TPU-timing” means that both BU and DU can be similarly timed. In the Mandarin utterance, BUs and DUs tend to have similar lengths when they contain the same number of tones, whereas in the Shanghai Wu one, they are less even in length. Therefore, the Shanghai Wu utterance is less TPU-timed than the Mandarin one. Note that in previous studies, Chinese rhythm is described in terms of syllable-timing. However, since the

CCL 2024

proposed structure is tone-based, the description of Chinese rhythm needs to change from syllable- to tone-based accordingly. This change happens to support the notion of "tone-based rhythm" proposed by Xu (2018), and using TPU- instead of syllable-timing is a step forward to realize this notion.

## 5 A comparison with other structures

Hirst (2005) adopted a simple syllable-based four-level structure to capture the melodic and rhythmic differences between English and French. For English, the four levels of units from the bottom up are syllable (Syl.), rhythm unit (RU), tonal unit (TU), and intonation unit (IU). An RU can cover a trochaic stress foot and monosyllables outside the foot, delimited by word boundaries. A TU can cover RUs, delimited either by a foot or by an IU boundary. This structure allows a nuclear accent to occur non-phrase-finally, reflecting the left-heavy prominence of English ([H L]). For French, the four levels of units are Syl., TU, RU, and IU, with RU and TU reversed. TU being placed under RU indicates that no nuclear accent occurs phrase-initially, reflecting the right-heavy prominence of French ([L H]). Accordingly, English has a falling melody while French a rising one, and English has a stress-timed rhythm while French a syllable-timed one.

The advantage of this structure is its ability to integrate prosodic functions into a hierarchy without immediate morpho-syntactic concerns. However, this simple structure is not applicable to Chinese, because Chinese syllables are mostly toned, and the prosodic and morpho-syntactic interface is mainly made visible through tone sandhi rather than stress foot. Also, TU and RU, even if applicable to Chinese, can be consolidated into one given the pitch-timing duality of tone, and the TPU is just such one.

In addition, the tone-based structure is applicable to various Chinese dialects and able to provide a unified interface between prosody and morpho-syntax. This is because the prosodic features of different Chinese dialects can be reflected by the nature, head position, and branching direction of their TPUs. As long as tone sandhi patterns are identified, two levels of TPU (BU/DU) can be formed regardless of whether they are associated with words, phrases, or clauses. Now English, French, and Chinese are compared below in terms of their four-level structures and prosodic patterns. A simplified conventional structure is also included.

| Prosodic structure | English | French | Chinese | Conventional |
|---|---|---|---|---|
| Level 4 | IU | IU | IU | IP |
| | ↑ | ↑ | ↑ | ↑ |
| Level 3 | TU | RU | DU | PP |
| | ↑ | ↑ | ↑ | ↑ |
| Level 2 | RU | TU | BU | PW |
| | ↑ | ↑ | ↑ | ↑ |
| Level 1 | Syl. | Syl. | PU | Syl. |
| Melodic pattern | [H L] | [L H] | [H L]/[L H] | Not Specified |
| Rhythmic pattern | Stress-timing | Syllable-timing | TPU-timing | Syllable-timing |

Figure 7. A Comparison of English, French, Chinese, and Conventional Four-level Structures.

The last two rows in Figure 7 suggests that a syllable-based structure, however simple, is neither fully nor directly able to reveal Chinese melodic and rhythmic patterns as the tone-based one does. This is because Chinese can have the melodic pattern either falling or rising (corresponding to left- or right-heavy) at

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1201-1211, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

different levels (hence [H L]/[L H]). Also, TPU-timing is a more appropriate term used to describe Chinese rhythm than stress- or syllable-timing, as it is entirely PU-based, and the PU number helps determine the tone and subsequent TPU length.

## 6 Implications of the proposed structure for prosody modeling

The above section shows that the proposed TP layer is related to different sizes or types of tone sandhi domains but does not directly refer to their morpho-syntactic status. In other words, TPU-based prosody processing does not need to rely heavily on grammatical markings, hence compatible with various theory-dependent models. However, any new model is subject to empirical validation, and the proposed one is no exception. In order to gain some insight for the future validation work, two prosody modeling studies are discussed below, because they are related to implications of the proposed structure.

First, Hirst (2015) developed a modeling tool, ProZed, based on the four-level English structure shown in Figure 7. This model is of theory-friendly nature and has an analysis-by-synthesis paradigm, hence allowing easy manipulation of rhythmic and tonal properties as well as immediate interactive experimentation with prosodic annotation. The proposed structure has a similar theory-friendly design, so it can be used to build a modeling tool to test different phonological theories on tonal behaviors. Better yet, it can accommodate a wide range of Chinese dialects with very different tone sandhi patterns and grammatical structures.

Second, Zu et al. (2022) discussed a so-called “sense element” (SE) in their experimental study of Chinese speech synthesis. The SE is a basic input unit encompassing both individual tones and multi-tone sandhi domains. Their results show that the SE-based input system leads to a higher MOS (Mean Opinion Score) than the traditional dictionary-based one in the assessment of the synthesized speech of Lhasa Tibetan. According to Zu et al. (2022), the main reason for SE to work well with Lhasa Tibetan data is its ability to maintain a good text and speech alignment, especially when functional words and verbal phrases are involved, both prone to phonology-syntax mismatch. When a mismatch occurs, the SE can be adjusted accordingly in size, since they are not confined to grammatical structures. The present study considers an SE in essence a BU or BU’, so it is expected that a BU-based system works as well. Nonetheless, the proposed structure is yet to be applied in future modeling projects and validated with favorable results.

## 7 Conclusion

The present study illustrates the tone-based prosodic structure and its application in two types of Chinese. Since this structure can accommodate both prosodically and morpho-syntatically based tone sandhi domains, there is no need to consider whether or not syllables are stressed: prosodic units are headed on each level and timed according to size. Unlike the above mentioned syllable-based structures, the proposed one not only helps capture melodic and rhythmic patterns of different Chinese dialects but also provides a theory-friendly framework for Chinese speech engineering.

CCL 2024

## References


Chengyu Guo. 2020. *Prosodic Structures and Tone Sandhi of Bi-syllabic Words in Chinese Dialects*. PhD Diss., Shanghai Normal University, Shanghai.

Daniel J. Hirst. 2005. Form and function in the representation of speech prosody. *Speech Communication,* 46(4): 334-347.

Daniel J. Hirst. 2013. Melody metrics for prosodic typology: comparing English, French, and Chinese. *Proceedings of Interspeech 2013*. 572–576.

Daniel J. Hirst. 2015. ProZed: A speech prosody editor for linguistic, using analysis-by-synthesis. In K Hirose, J Tao (eds.), *Speech Prosody in Speech Synthesis: Modeling and Generation of Prosody for High Quality and Flexible Speech Synthesis*. Springer Verlag, Berlin Heidelberg. Chapter 1: 3–17.

Guhui Li. 2017. *On the Syntax-prosody Interface: Evidence from Tone Sandhi Processes in Chinese*. PhD Diss., East China Normal University, Shanghai.

Hongming Zhang. 2017. *Syntax-Phonology Interface: Argumentation from Tone Sandhi in Chinese Dialects*. Routledge, New York.

Huibin Zhuang. 2015. On defining *ci* in Chinese: A perspective from prosodic grammar. *TCSOL Studies*, 58(2):61–69.

Jianfen Cao. 2011. Prosodic structure and speech sound variation. *Journal of School of Chinese Language and Culture, Nanjing Normal University,* 3: 12–22.

Marina Nespor and Irene Vogel. 1986. *Prosodic Phonology*. Foris, Dordrecht, Netherlands.

Matthew Y. Chen. 2000. *Tone Sandhi: Patterns across Chinese Dialects*. Cambridge University Press, Cambridge, MA.

Peter Ladefoged and Keith Johnson. 2011. *A Course in Phonetics (Sixth Edition)*. Wadsworth, Boston, MA.

Qiuwu, Ma. 2021. Chinese word stress and its phonological properties. *Studies in Prosodic Grammar*, 8: 112–129.

Xi Ming Xu. 2018. *Contrastive Studies of Rhythmic Types in English and Mandarin.* Foreign Language Teaching and Research Press, Beijing.

Xiaobin Yan. 2010. A critical review of research on the phrasing of Chinese tone sandhi. *Journal of Nanjing University of Science and Technology (Social Sciences)*, 23(4): 76–83.

Ya Li. 2015. *Timing and Melody: An Acoustic Study of Rhythmic Patterns of Chinese Dialects.* PhD Diss., University of Victoria, Victoria, Canada.

Ya Li. 2023. Timing specification of Chinese tones. *International Journal of Chinese Linguistics*, 11(1): 76–97.

Yiqing Zu, Chen Lu, Ngodrup, Ronghua Zhu, Chenning Liu, Pengfei Shao, Klu' Bum Thr, Xiao Zhang, and Guoping Hu. 2022. Sense element in continuous speech: Evidence from Lhasa Tibetan speech synthesis. *Contemporary Linguistics*, 24(4): 515–532.

Youwei Shi. 2017. The dual-unit system in Chinese: Thoughts on the units of Chinese language teaching. *Journal of International Chinese Teaching*, 14(2): 34–41.

Zhigang Yin. 2020. Revisiting the acoustic characteristics and the generation mechanism of prosodic boundary. *Chinese Journal of Phonetics*, 13: 38–50.


CCL 2024

# Linguistic Guidance for Sequence-to-Sequence AMR Parsing

**Binghao Tang, Boda Lin, Si Li***
School of Artificial Intelligence, Beijing University of Posts and Telecommunications
{tangbinghao,linboda,lisi}@bupt.edu.cn

## Abstract

The Abstract Meaning Representation (AMR) parsing aims at capturing the meaning of a sentence in the form of an AMR graph. Sequence-to-sequence (seq2seq)-based methods, utilizing powerful Encoder-Decoder pre-trained language models (PLMs), have shown promising performance. Subsequent works have further improved the utilization of AMR graph information for seq2seq models. However, seq2seq models generate output sequence incrementally, and inaccurate subsequence at the beginning can negatively impact final outputs, also the interconnection between other linguistic representation formats and AMR remains an underexplored domain in existing research. To mitigate the issue of error propagation and to investigate the guiding influence of other representation formats on PLMs, we propose a novel approach of **L**inguistic **G**uidance for **S**eq2seq **A**MR parsing (**LGSA**). Our proposed LGSA incorporates the very limited information of various linguistic representation formats as guidance on the Encoder side, which can effectively enhance PLMs to their further potential, and boost AMR parsing. The results on proverbial benchmark AMR2.0 and AMR3.0 demonstrate the efficacy of LGSA, which can improve seq2seq AMR parsers without silver AMR data or alignment information. Moreover, we evaluate the generalization of LGSA by conducting experiments on out-of-domain datasets, and the results indicate that LGSA is even effective in such challenging scenarios.

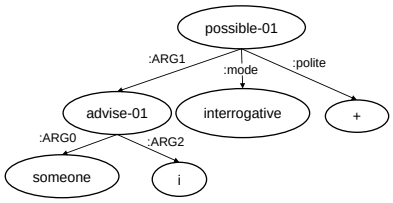
## 1 Introduction

Abstract Meaning Representation (AMR) (Banarescu et al., 2013) is a semantic representation framework that systematically converts the meaning of a sentence into a rooted, labeled, and directed acyclic graph (DAG), where the nodes represent concepts and the edges denote semantic relations. Figure 1 depicts an example of an AMR graph. AMR has gained significant traction in various downstream natural language processing (NLP) tasks, including Information Extraction (Rao et al., 2017; Huang et al., 2016; Zhang and Ji, 2021), Text Summarization (Hardy and Vlachos, 2018; Liao et al., 2018), Machine Translation (Song et al., 2019), and Dialog System (Bonial et al., 2020; Bai et al., 2021).

The methods of AMR parsing can be divided into four categories: two-stage parsing (Flanigan et al., 2014; Flanigan et al., 2016; Lyu and Titov, 2018; Zhang et al., 2019a), transition-based parsing (Wang et al., 2015; Damonte et al., 2017; Vilares and Gómez-Rodríguez, 2018; Liu et al., 2018; Naseem et al., 2019; Fernandez Astudillo et al., 2020; Zhou et al., 2021a; Zhou et al., 2021b), graph-based parsing (Zhang et al., 2019b; Cai and Lam, 2019; Cai and Lam, 2020) and sequence-to-sequence (seq2seq)-based parsing (Barzdins and Gosko, 2016; Konstas et al., 2017; Peng et al., 2017; Ge et al., 2019; Bevilacqua et al., 2021; Yu and Gildea, 2022; Bai et al., 2022). In contrast to other approaches, seq2seq-based methods view AMR parsing as the process of transforming the input sentence into a linearized sequence of AMR graph. This allows for the seamless integration of powerful pre-trained seq2seq Transformer models, which have shown significant advancements in various NLP tasks. As a result,

*Corresponding author
©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1212-1222, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China




**Gold DFS Linearization:**

(<R0> possible-01 :ARG1 (<R1> advise-01
(:ARG0 (<R2> someone) :ARG2 (<R3>
i))) :mode (<R4> interrogative) :polite <R5> +)

**Pred DFS Linearization:**

(<R0> advise-01 :ARG0 (<R1>
someone) :ARG1 (<R2> i) :mode (<R3>
interrogative) :polite <R4> +)

Figure 1: An AMR graph for sentence "Can someone give me advice please?" with its corresponding gold and predicted DFS-based linearized sequence. The prediction has the wrong root token and leads to a worse AMR graph compared with the gold AMR graph.

seq2seq-based methods have achieved remarkable performance, positioning them at the forefront of current research in AMR parsing.

Seq2seq-based works employ various techniques and auxiliary features while the performance is still unsatisfactory. For instance, structure-aware encoding (Ge et al., 2019) explicitly models syntax and semantics through complex source sequence linearization of the whole syntactic tree with semantic roles based on pure seq2seq Encoder-Decoder Transformer. More recently, seq2seq-based methods applying Pre-trained Language Models (PLMs) like BART (Lewis et al., 2020), further improve seq2seq-based methods by a large margin. These approaches focus on enhancing AMR parsers from various perspectives, such as AMR graph linearization (Bevilacqua et al., 2021), revised Decoder Attention (Yu and Gildea, 2022), curriculum learning (Wang et al., 2022), intermediate-task learning (Chen et al., 2022), graph pre-training (Bai et al., 2022), word-to-node alignment (Vasylenko et al., 2023) and instruction-tuning (Lee et al., 2023). What becomes apparent is that the methods before the adoption of encoder-decoder PLMs involved complex linearization and the utilization of the other whole linguistic representation formats. They also incorporated a significant amount of structural information that might not have been necessary for the AMR task, which could instead result in noise.

The subsequent approaches shifted their focus towards better harnessing the inherent structural information and alignments within the AMR data based on encoder-decoder pre-trained language models. However, the output AMR graph of seq2seq parsers is contingent upon the previously generated sub-graph, so the inaccurate sub-graph may lead to error propagation. Building upon the above issues, we aim to demonstrate that using the very limited information of different linguistic formats as guidance could still prompt the PLMs to generate better output tokens and reduce error propagation in AMR parsers, further improving their performance.

In this paper, we introduce a novel model-agnostic approach of **L**inguistic **G**uidance for **S**eq2seq **A**MR parsing (**LGSA**). The proposed LGSA leverages the very limited other linguistic representations as guidance on the encoder side and prompts the seq2seq-based parsers to generate more accurate sub-graphs. Specifically, we choose three commonly used linguistic representation formats: Part-of-Speech (POS) that can represent lexical features, Dependency Parsing (DP) that can represent syntactic features, and Semantic Dependency Parsing (SDP) (Oepen et al., 2015) that can represent semantic features. We build LGSA on the top of two recent seq2seq AMR parsers: SPRING (Bevilacqua et al., 2021) and Ancestors (Yu and Gildea, 2022). Experimental results on proverbial benchmark AMR2.0 and AMR3.0 demonstrate our proposed LGSA can further enhance the ability of the seq2seq AMR parsers. Also, Ancestors with LGSA of SDP guidance even achieves the comparable performance of AMRBART (Bai et al., 2022), which utilizes 200k silver AMR data for the post graph pre-training based on BART. We also conduct out-of-domain (OOD) experiments on three OOD datasets New3, TLP, and Bio, which confirm the effectiveness of the proposed LGSA method even in OOD situations.

## 2 Related Work

**AMR Parsing** AMR Parsing is a fundamental task that transforms a given sentence into a corresponding AMR graph. The methods of AMR parsing can be divided into four categories: two-stage parsing,

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1212-1222, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

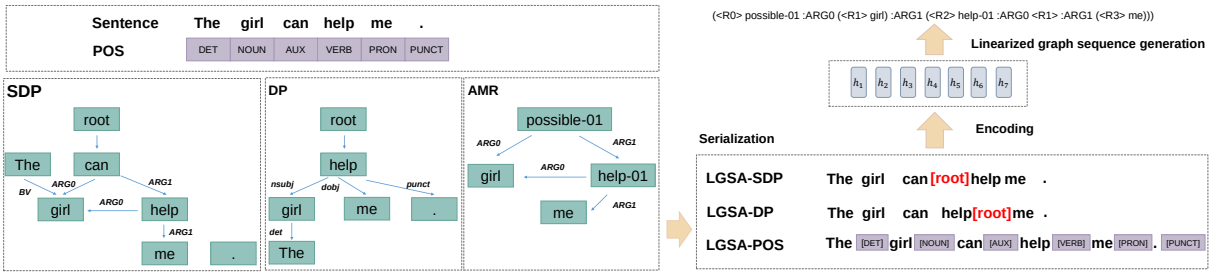



Figure 2: Serialization of our proposed LGSA, which utilizes the limited information of Semantic Dependency Parsing (SDP), Dependency Parsing (DP), and Part-of-Speech (POS) as the linguistic guidance to prompt PLMs. The obtained linguistic guidance source sequence is input to sequence-to-sequence AMR parsers to generate the corresponding linearized AMR graph.

transition-based parsing, graph-based parsing, and seq2seq-based parsing. Two-stage parsing first predicts potential concepts nodes, and then labels corresponding relations based on previous concepts decisions (Flanigan et al., 2014; Flanigan et al., 2016; Lyu and Titov, 2018; Zhang et al., 2019a). Transition-based methods (Wang et al., 2015; Fernandez Astudillo et al., 2020; Zhou et al., 2021a; Zhou et al., 2021b) process tokens of the input sentence from left to right and generate graph incrementally based on a pre-defined set of actions. Graph-based methods (Lyu and Titov, 2018; Zhang et al., 2019a; Cai and Lam, 2020) take AMR parsing as an incremental sequence-to-graph transduction task. Seq2seq-based methods treat AMR parsing as a sequence-to-sequence problem by AMR graph linearization. Several techniques have been employed in the early stages of many seq2seq works to improve the performance, such as auxiliary features (Konstas et al., 2017). For instance, (Ge et al., 2019) explicitly modeled syntax and semantics through linearized source sequence based on a syntactic tree with semantic roles and revised structure-aware self-attention. However, the performance of seq2seq-based methods has remained unsatisfactory until the introduction of SPRING (Bevilacqua et al., 2021). SPRING presented a better AMR graph linearization procedure and utilized BART (Lewis et al., 2020) as the backbone, which led to a significant improvement in seq2seq-based AMR Parsing. Subsequent seq2seq-based works have primarily focused on fully utilizing the structural information of AMR, such as adding ancestors information to Transformer Decoder (Yu and Gildea, 2022), utilizing different training methods like curriculum learning (Wang et al., 2022) and using extra silver data for intermediate-task learning (Chen et al., 2022) or further pretraining (Bai et al., 2022). In contrast, we pay attention to utilizing the very limited information of other linguistic formats to guide PLMs in a model-agnostic way.

**Dependency Parsing and Semantic Dependency Parsing** Dependency Parsing (DP) is a fundamental task in Natural Language Processing, aiming at representing the sentence in a format of tree structure. DP hopes to identify the syntactic relationships between words in a sentence. So DP is typically as syntactic features. The 2014 SemEval (Oepen et al., 2014; Oepen et al., 2015) shared task introduced a new representation of the sentence which focuses on the deep-semantic relations between words. Semantic dependency parsing involves converting an input sentence into a directed acyclic graph called a Semantic Dependency Graph (SDG). Due to the similarity between SDG and AMR graphs, which are both directed acyclic graphs, the structural information of SDG can serve as a supplement to AMR.

## 3 Methodology

### 3.1 Task Formulation

For seq2seq-based methods, the AMR parsing task is usually formulated as a transformation of the input sentence and linearized AMR sequence. Given input sentence $S = \{t_1, t_2, ..., t_n\}$ , where $t_i$ is the $i$-th token of the input sentence, model generates linearized AMR graph sequence $Y = \{y_1, y_2, ..., y_n\}$ , where $y_i$ is the $i$-th token of the AMR linearized sequence. As Figure 1 shows, the DFS-based linearization of SPRING (Bevilacqua et al., 2021) which uses special tokens like $\{\langle R_0\rangle, ..., \langle R_n\rangle\}$ to represent

variables in the linearized graph, has been widely applied in seq2seq-based methods. However, in the case of the seq2seq model, the probability of generating the next token at the decoder is contingent upon the conditional probability of all previously generated tokens: $P(y_{i+1}|y_1, \ldots, y_i)$. Consequently, the accuracy of the initial token significantly influences the overall outcome. In this paper, we propose a novel model-agnostic approach of linguistic guidance for seq2seq AMR parsing, which provides the model with a modest amount of information on other linguistic representation formats as prompts to improve final AMR outcomes. We feed our linguistic serialized results into the Transformer Encoder-Decoder pre-trained language model instead of the original sentence. Then the model generates a corresponding AMR graph linearized sequence.

### 3.2 Linguistic Guidance Serialization

Serialization is a common technique used in seq2seq tasks, where various serializers have been designed for syntactic and semantic parsing (Ge et al., 2019; Lin et al., 2022). These serializers generate a serialized representation based on the specific linguistic representation, aiming at improving the performance of parsers.

Taking inspiration from these works, we hope to seek a straightforward approach to leverage a limited amount of other linguistic representation formats to guide seq2seq AMR parsers, which can be divided into parsing and lexical information[0]:

**Parsing Information** SDP and DP are selected as semantic and syntactic parsing information. SDP emphasizes semantics and converts the sentence into a DAG while DP focuses on dependency relations between words and represents the sentence in a format of tree structure. As shown in Figure 2, the input sentence "The girl can help me." has three forms of representation, SDP, DP, and AMR. The focal concept node of AMR form is *possible-01*, which corresponds to *can* in the original sentence. In the form of SDP, the root token of the sentence is *can* as well. This observation suggests that the root token of SDP has a natural alignment with the focal concept in AMR. Additionally, both SDP root tokens and AMR focal concepts are typically related to verbs, which serve a predicative function in sentences. Similarly, the syntactic root tokens of DP are also frequently linked to verbs like SDP and AMR. Thus we introduce a special token *[root]* to highlight the SDP and DP root tokens in the original sentences. We append this token to the respective root tokens and obtain LGSA-SDP and LGSA-DP sequences, as shown in Figure 2. This is done in the hope of choosing SDP and DP root tokens as soft guidance to prompt AMR parsers to generate more accurate tokens. Furthermore, based on the above procedures, we can simply serialize the source sequence using the root information of SDP and DP, without the need to employ a complete parsing tree for complex serialization or introduce numerous unnecessary parsing labels.

**Lexical Information** For lexical information, we select POS tagging, which is demonstrated beneficial for both semantic and syntactic parsing (Zhang et al., 2019a). To integrate the POS information into our serialization, we simply append the corresponding POS tag to the original token. Then the POS tag embedding can be added to the original sentence at the outset, as shown in Figure 2. To point out, compared to parsing information, we believe that lexical information may not be significantly helpful for AMR. Furthermore, POS information is not as easily utilized partially as parsing information. Therefore, we decided to fully utilize the entire POS information for comparison, to validate our aforementioned hypothesis.

In the aforementioned LGSA method, we can utilize a certain amount of external linguistic knowledge to preprocess the input sequence, highlighting the key information, and leveraging the powerful capabilities of PLMs to learn better AMR graph prediction.

[0]Different from (Ge et al., 2019), we aim to improve AMR parsers with limited utilization of extra-linguistic information based on PLMs.

| Datasets | **Train** | **Dev** | **Test** |
|---|---|---|---|
| AMR2.0 | 36,521 | 1,371 | 1,368 |
| AMR3.0 | 55,635 | 1,722 | 1,898 |
| New3 | 4,441 | 354 | 527 |
| TLP | – | – | 1,562 |
| Bio | 5,452 | – | 500 |

Table 1: Datasets statistics

## 4 Experiments

### 4.1 Datasets and Metrics

We evaluate our method on two proverbial AMR benchmarks, AMR2.0 (LDC2017T10) and AMR3.0 (LDC2020T02). Investigating OOD performance, we follow the same OOD experiment setting of SPRING (Bevilacqua et al., 2021), employ AMR2.0 training set, and evaluate on three OOD datasets, New3, Bio, and TLP. New3 is a subset of AMR3.0 but not contained in AMR2.0, which has 527 sentences. TLP is an AMR-annotated novel "The Little Prince" consisting of 1562 sentences. Bio is from the biomedical AMR corpus and has 500 sentences. The statistics are shown in Table 1.

We assess the performance of our parsers with Smatch (Cai and Knight, 2013) scores computed by amrlib [1] tools, which also provide fine-grained scores:

- Unlabeled: Smatch score after removing edge-labels
- NoWSD: Smatch score after ignoring Prop-bank senses (e.g., come-01 vs come-02)
- Concepts: F-score on the concept identification task
- Wikification (Wiki.): F-score on the wikification (:wiki roles)
- Named Entity Recognition (NER): F-score on the named entities (:name roles).
- Reentrancy (Reent.): Smatch score on reentrant edges.
- Negation (Neg.): F-score on the negation detection (:polarity roles)
- Semantic Role Labeling (SRL): Smatch score computed on :ARG-i roles only.

The fine-grained scores assess the predictions and gold AMR graphs from different angles.

### 4.2 Experiment Setting

For linguistic serialization, we choose Stanford NLP tools [2] to do tokenization and POS tagging. We use the existing parsing pipeline SuPar [3] to obtain the DM (DELPH-IN MRS) representation of the semantic dependency graph and dependency tree of the raw input sentence. We use BART-large (Lewis et al., 2020) as seq2seq pre-trained model backbone and RAdam (Liu et al., 2019) as optimizer. The learning rate is 3e-5, batch size is 1000, and beam size is 5. The epochs are 50 and the loss function is cross-entropy. The best model is selected according to performance on the validation set. All models are trained and evaluated on a single NVIDIA 3090Ti GPU. We conduct the same pre-processing and post-processing of the linearized AMR graph following SPRING (Bevilacqua et al., 2021). However, we do not adopt graph recategorization, which is usually used to alleviate data sparsity in graph-based methods yet has been proved to be harmful for OOD situations (Bevilacqua et al., 2021).

CCL 2024

| Model | Extra Data | Smatch | No WSD | Concept | NER | Neg. | Wiki. | Unlabeled | Reent. | SRL |
|---|---|---|---|---|---|---|---|---|---|---|
| **AMR2.0** | | | | | | | | | | |
| HCL (Wang et al., 2022) | N | 84.3 | 85.0 | 90.2 | 91.6 | **75.9** | 84.0 | 87.7 | 74.5 | 83.2 |
| ATP(w/SRL) (Chen et al., 2022) | 40k | 85.2 | 85.6 | 90.7 | **93.1** | 74.9 | **84.2** | 88.3 | **74.7** | **83.3** |
| AMRBART (Bai et al., 2022) | 200k | **85.4** | **85.8** | **91.2** | 91.5 | 74.0 | 81.4 | **88.3** | 73.5 | 81.5 |
| SPRING (Bevilacqua et al., 2021) | N | 83.8 | 84.4 | 90.2 | 90.6 | 74.4 | **84.3** | 86.1 | 70.8 | 79.6 |
| SPRING(w/silver) (Bevilacqua et al., 2021) | 200k | 84.3 | 84.8 | 90.8 | 90.5 | 73.6 | 83.1 | 86.7 | 72.4 | 80.5 |
| SPRING(w/LGSA-POS) | 50k | 84.2 | 84.8 | 90.7 | 90.8 | 74.6 | 83.9 | 86.4 | 71.9 | 80.1 |
| SPRING(w/LGSA-DP) | 50k | 84.4 | 84.8 | 90.7 | 90.9 | 75.8 | 84.0 | 86.6 | 71.9 | 80.3 |
| SPRING(w/LGSA-SDP) | 50k | **84.5** | **85.0** | 90.7 | **91.3** | **76.2** | 84.1 | **86.8** | **72.4** | **80.5** |
| Ancestors (Yu and Gildea, 2022) | N | 84.8 | 85.3 | 90.5 | **91.8** | 74.0 | 84.1 | 88.1 | **75.1** | **83.4** |
| Ancestors(w/LGSA-POS) | 50k | 85.1 | 85.6 | 90.9 | 90.6 | **75.1** | **84.7** | 88.0 | 73.7 | 81.1 |
| Ancestors(w/LGSA-DP) | 50k | 85.2 | 85.6 | 90.9 | 91.4 | 73.8 | 83.5 | 88.1 | 73.7 | 81.4 |
| Ancestors(w/LGSA-SDP) | 50k | **85.4** | **85.8** | **91.1** | 91.3 | 74.6 | 83.9 | **88.2** | 74.1 | 81.5 |
| **AMR3.0** | | | | | | | | | | |
| HCL (Wang et al., 2022) | N | 83.7 | 84.2 | 89.5 | **89.0** | 73.0 | **82.6** | 86.9 | 73.9 | 82.4 |
| ATP(w/SRL) (Chen et al., 2022) | 40k | 83.9 | 84.3 | 89.7 | 88.4 | **73.9** | 81.0 | 87.0 | **73.9** | **82.5** |
| AMRBART (Bai et al., 2022) | 200k | **84.2** | **84.6** | **90.2** | 88.5 | 72.1 | 78.9 | **87.1** | 72.4 | 80.3 |
| SPRING (Bevilacqua et al., 2021) | N | 83.0 | 83.5 | 89.8 | 87.2 | **73.0** | **82.7** | 85.4 | 70.4 | 78.9 |
| SPRING(w/silver) (Bevilacqua et al., 2021) | 200k | 83.0 | 83.5 | 89.5 | 87.1 | 71.7 | 81.2 | 85.4 | 71.3 | 79.1 |
| SPRING(w/LGSA-POS) | 50k | 83.5 | 83.8 | 89.8 | 87.1 | 72.4 | 81.9 | 85.9 | 71.0 | 78.7 |
| SPRING(w/LGSA-DP) | 50k | 83.6 | 84.0 | 91.1 | 87.2 | 72.6 | 82.1 | 86.1 | 71.2 | 78.9 |
| SPRING(w/LGSA-SDP) | 50k | **83.8** | **84.1** | **91.3** | **88.0** | 72.9 | 82.3 | **86.3** | **71.4** | **79.1** |
| Ancestors (Yu and Gildea, 2022) | N | 83.5 | 84.0 | 89.5 | 88.9 | 72.6 | 81.5 | 86.6 | **74.2** | **82.2** |
| Ancestors(w/LGSA-POS) | 50k | 83.9 | 84.4 | 90.3 | 88.7 | 71.9 | 82.0 | 86.9 | 72.5 | 80.1 |
| Ancestors(w/LGSA-DP) | 50k | 84.0 | 84.4 | 90.2 | 88.4 | 73.1 | 81.7 | 86.9 | 72.7 | 80.2 |
| Ancestors(w/LGSA-SDP) | 50k | **84.2** | **84.6** | **90.5** | **89.0** | **73.1** | **82.2** | **87.2** | 72.7 | 80.4 |

Table 2: Experimental results on AMR2.0 and AMR3.0, our proposed LGSA does not fully leverage 50k extra data but rather only utilizes partial information as guidance to prompt PLMs.

### 4.3 Main results

We build our method on the top of seq2seq AMR parsers: SPRING (Bevilacqua et al., 2021) and Ancestors (Yu and Gildea, 2022), and show the improvement of our method. We also present the experimental results of recent seq2seq-based methods. To point out, the compared methods ATP (Chen et al., 2022) with Semantic Role Labeling (SRL) uses 40k SRL data from OntoNotes (LDC2013T19) for intermediate-task learning and AMRBART (Bai et al., 2022) leverage 200k silver data from Gigaword (LDC2011T07) parsed by SPRING for further graph pretraining on the basis of vanilla BART while our proposed LGSA method obtain extra-linguistic information only based on the original sentences and do not require any additional corpora.

The main results are as shown in Table 2. Among the three types of linguistic information, LGSA-SDP shows the highest improvement and LGSA-POS performs the worst. More precisely, SPRING with LGSA-SDP achieves an improvement of 0.7 Smatch on AMR2.0 and 0.8 Smatch on AMR3.0, and Ancestors with LGSA-SDP achieves an improvement of 0.6 Smatch on AMR2.0 and 0.7 Smatch on AMR3.0 compared with baseline models. The results suggest that incorporating semantic parsing information (SDP) is more helpful for AMR parsing than incorporating lexical information (POS). This is likely because AMR is a semantic representation, therefore, semantic features should be more relevant to capturing its meaning. Furthermore, the results show that Ancestors with LGSA-SDP achieves competitive performance even against models that utilize a large amount of silver AMR data such as AMRBART (Bai et al., 2022) and ATP (Chen et al., 2022). This indicates that our approach can achieve promising results without requiring any additional corpora, but simply by leveraging the linguistic features extracted from the original sentences and utilize a limited amount of information to be guidance on the encoder side. For fine-grained scores, Ancestors with LGSA-SDP also obtain the best result of No

[1] https://github.com/bjascob/amrlib
[2] https://github.com/stanfordnlp/stanza
[3] https://github.com/yzhangcs/parser

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1212-1222, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

| Model | New3 | TLP | Bio |
|---|---|---|---|
| SPRING (Bevilacqua et al., 2021) | 73.7 | 77.3 | 59.7 |
| SPRING(w/LGSA-POS) | 73.9 | 77.5 | 61.0 |
| SPRING(w/LGSA-DP) | 74.0 | 77.5 | 61.2 |
| SPRING(w/LGSA-SDP) | **74.1** | **77.7** | **61.4** |
| Ancestors (Yu and Gildea, 2022) | 75.2 | 79.3 | 61.4 |
| Ancestors(w/LGSA-POS) | 75.3 | 79.3 | 61.6 |
| Ancestors(w/LGSA-DP) | 75.4 | 79.4 | 61.8 |
| Ancestors(w/LGSA-SDP) | **75.6** | **79.4** | **62.0** |

Table 3: Experimental results on OOD datasets.

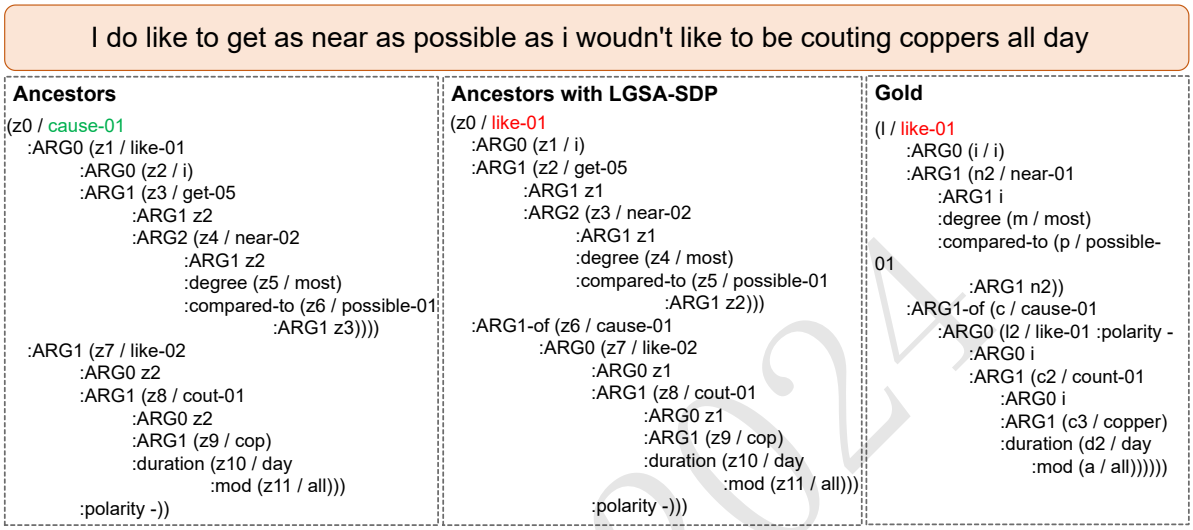



Figure 3: A case selected from AMR2.0 test set. For the input sentence, the baseline model Ancestors with LGSA-SDP achieves the right root token and a more accurate AMR graph, while Ancestors (Yu and Gildea, 2022) gets a relatively worse result. The results show that using Ancestors alone yields a Smatch score of 63.2 while incorporating LGSA-SDP improves it to 73.7.

WSD on AMR2.0 and No WSD, Concept, NER and Unlabeled on AMR3.0.

The main experimental results prove that our proposed LGSA method can enhance the performance of the current seq2seq AMR parsers by utilizing a limited amount of external linguistic knowledge, which aligns with our initial intention stated in the introduction. It also demonstrates that linguistic representations are an interconnected whole, where different forms of representation mutually benefit each other. The closer the forms of representation, the more pronounced this beneficial effect, as exemplified by SDP and AMR.

### 4.4 OOD Results

Previous methods have demonstrated promising results on in-domain datasets, but their performance significantly declines on OOD datasets, highlighting the importance of generalization in such methods. In this paper, we compare our proposed LGSA method with our baseline seq2seq parsers. As illustrated in Table 3, LGSA-SDP still achieves the best results among all the methods on three OOD datasets whether it is SPRING or Ancestors. Also, LGSA-DP performs worse and LGSA-POS performs the worst, which is consistent with their performance on the main in-domain dataset. These results demonstrate the efficacy of our proposed LGSA method in utilizing other linguistic representation formats to guide AMR parsers and enhance performance in OOD scenarios. Yet we believe that the limited performance improvement on OOD datasets is due to the accuracy of the parsing pipeline on OOD data. Given the use of very limited information, the accuracy like root tokens limits the performance of AMR parsers in

CCL 2024

OOD scenarios. Despite these challenges, our proposed LGSA method has still proven effective in OOD scenarios across multiple runs of experiments.

### 4.5 Case Study

We further conduct a case study on the seq2seq parser Ancestors (Yu and Gildea, 2022) to explore the output results with and without incorporating our method, thereby proving the effectiveness of our approach. Given that the LGSA-SDP shows the best improvement on both main and OOD results, we compare the output AMR graphs of the baseline model Ancestors (Yu and Gildea, 2022) and Ancestors (Yu and Gildea, 2022) with LGSA-SDP.

As is shown in Figure 3, Ancestors (Yu and Gildea, 2022) with LGSA-SDP gets the right focus concept *like-01* with the aid of SDP, and generates a more accurate AMR graph while Ancestors gets the wrong focus concept *cause-01* and results in less correct structure. The baseline prediction gets Smatch 63.2 computed with gold graph while ours obtains Smatch 73.7. This indicates that with SDP as an additional guidance, compared to the original output, the seq2seq AMR parser Ancestors can better predict the correct root token. This leads to a more accurate prediction of the corresponding AMR graph, thereby enhancing the performance of AMR parsers. In this particular sample case, with only the very limited additional guidance provided like the SDP root token, the model shows a ten-point improvement in the Smatch score. This also demonstrates that for AMR parsers based on PLMs, the use of limited external linguistic knowledge can significantly enhance their current performance.

## 5 Conclusions

In this paper, we focus on the error propagation of seq2seq parsers and investigate to utilize a limited amount information of from other linguistic representation formats as guidance for seq2seq AMR parsing to better utilize pre-trained language models. We propose a novel model-agnostic approach called **L**inguistic **G**uidance for **S**eq2seq **A**MR Parsing (**LGSA**), which leverages Part-of-Speech (POS) information for lexical features, Dependency Parsing for syntactic features, and Semantic Dependency Parsing for semantic features. Experiments on AMR2.0 and AMR3.0 demonstrate that LGSA can incorporate additional linguistic information on the Encoder side and enhance the ability of AMR parsers. Furthermore, we also demonstrate the effectiveness of LGSA in out-of-domain (OOD) benchmarks including New3, TLP, and Bio.

## Acknowledgements

This work was supported in part by National Natural Science Foundation of China (NSFC) No. U23B2052, Program for Youth Innovative Research Team of BUPT No. 2023QNTD02.

## References

Xuefeng Bai, Yulong Chen, Linfeng Song, and Yue Zhang. 2021. Semantic representation for dialogue modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4430–4445, Online, August. Association for Computational Linguistics.

Xuefeng Bai, Yulong Chen, and Yue Zhang. 2022. Graph pre-training for AMR parsing and generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6001–6015, Dublin, Ireland, May. Association for Computational Linguistics.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria, August. Association for Computational Linguistics.

Guntis Barzdins and Didzis Gosko. 2016. RIGA at SemEval-2016 task 8: Impact of Smatch extensions and character-level neural translation on AMR parsing accuracy. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1143–1147, San Diego, California, June. Association for Computational Linguistics.

Michele Bevilacqua, Rexhina Blloshmi, and Roberto Navigli. 2021. One spring to rule them both: Symmetric amr semantic parsing and generation without a complex pipeline.

Claire Bonial, Lucia Donatelli, Mitchell Abrams, Stephanie M. Lukin, Stephen Tratz, Matthew Marge, Ron Artstein, David Traum, and Clare Voss. 2020. Dialogue-AMR: Abstract Meaning Representation for dialogue. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 684–695, Marseille, France, May. European Language Resources Association.

Shu Cai and Kevin Knight. 2013. Smatch: an evaluation metric for semantic feature structures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 748–752, Sofia, Bulgaria, August. Association for Computational Linguistics.

Deng Cai and Wai Lam. 2019. Core semantic first: A top-down approach for AMR parsing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3799–3809, Hong Kong, China, November. Association for Computational Linguistics.

Deng Cai and Wai Lam. 2020. AMR parsing via graph-sequence iterative inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1290–1301, Online, July. Association for Computational Linguistics.

Liang Chen, Peiyi Wang, Runxin Xu, Tianyu Liu, Zhifang Sui, and Baobao Chang. 2022. ATP: AMRize then parse! enhancing AMR parsing with PseudoAMRs. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2482–2496, Seattle, United States, July. Association for Computational Linguistics.

Marco Damonte, Shay B. Cohen, and Giorgio Satta. 2017. An incremental parser for Abstract Meaning Representation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 536–546, Valencia, Spain, April. Association for Computational Linguistics.

Ramón Fernandez Astudillo, Miguel Ballesteros, Tahira Naseem, Austin Blodgett, and Radu Florian. 2020. Transition-based parsing with stack-transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1001–1007, Online, November. Association for Computational Linguistics.

Jeffrey Flanigan, Sam Thomson, Jaime Carbonell, Chris Dyer, and Noah A. Smith. 2014. A discriminative graph-based parser for the Abstract Meaning Representation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1426–1436, Baltimore, Maryland, June. Association for Computational Linguistics.

Jeffrey Flanigan, Chris Dyer, Noah A. Smith, and Jaime Carbonell. 2016. CMU at SemEval-2016 task 8: Graph-based AMR parsing with infinite ramp loss. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1202–1206, San Diego, California, June. Association for Computational Linguistics.

DongLai Ge, Junhui Li, Muhua Zhu, and Shoushan Li. 2019. Modeling source syntax and semantics for neural amr parsing. In *IJCAI*, pages 4975–4981.

Hardy Hardy and Andreas Vlachos. 2018. Guided neural language generation for abstractive summarization using Abstract Meaning Representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 768–773, Brussels, Belgium, October-November. Association for Computational Linguistics.

Lifu Huang, Taylor Cassidy, Xiaocheng Feng, Heng Ji, Clare R. Voss, Jiawei Han, and Avirup Sil. 2016. Liberal event extraction and event schema induction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 258–268, Berlin, Germany, August. Association for Computational Linguistics.

Ioannis Konstas, Srinivasan Iyer, Mark Yatskar, Yejin Choi, and Luke Zettlemoyer. 2017. Neural AMR: Sequence-to-sequence models for parsing and generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 146–157, Vancouver, Canada, July. Association for Computational Linguistics.

Young-Suk Lee, Ramón Fernandez Astudillo, Radu Florian, Tahira Naseem, and Salim Roukos. 2023. Amr parsing with instruction fine-tuned pre-trained language models. *arXiv preprint arXiv:2304.12272*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July. Association for Computational Linguistics.

Kexin Liao, Logan Lebanoff, and Fei Liu. 2018. Abstract Meaning Representation for multi-document summarization. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1178–1190, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Boda Lin, Zijun Yao, Jiaxin Shi, Shulin Cao, Binghao Tang, Si Li, Yong Luo, Juanzi Li, and Lei Hou. 2022. Dependency parsing via sequence generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7339–7353, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Yijia Liu, Wanxiang Che, Bo Zheng, Bing Qin, and Ting Liu. 2018. An AMR aligner tuned by transition-based parser. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2422–2430, Brussels, Belgium, October-November. Association for Computational Linguistics.

Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. 2019. On the variance of the adaptive learning rate and beyond. In *International Conference on Learning Representations*.

Chunchuan Lyu and Ivan Titov. 2018. AMR parsing as graph prediction with latent alignment. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 397–407, Melbourne, Australia, July. Association for Computational Linguistics.

Tahira Naseem, Abhishek Shah, Hui Wan, Radu Florian, Salim Roukos, and Miguel Ballesteros. 2019. Rewarding Smatch: Transition-based AMR parsing with reinforcement learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4586–4592, Florence, Italy, July. Association for Computational Linguistics.

Stephan Oepen, Marco Kuhlmann, Yusuke Miyao, Daniel Zeman, Dan Flickinger, Jan Hajič, Angelina Ivanova, and Yi Zhang. 2014. SemEval 2014 task 8: Broad-coverage semantic dependency parsing. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 63–72, Dublin, Ireland, August. Association for Computational Linguistics.

Stephan Oepen, Marco Kuhlmann, Yusuke Miyao, Daniel Zeman, Silvie Cinková, Dan Flickinger, Jan Hajič, and Zdeňka Urešová. 2015. SemEval 2015 task 18: Broad-coverage semantic dependency parsing. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 915–926, Denver, Colorado, June. Association for Computational Linguistics.

Xiaochang Peng, Chuan Wang, Daniel Gildea, and Nianwen Xue. 2017. Addressing the data sparsity issue in neural AMR parsing. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 366–375, Valencia, Spain, April. Association for Computational Linguistics.

Sudha Rao, Daniel Marcu, Kevin Knight, and Hal Daumé III. 2017. Biomedical event extraction using Abstract Meaning Representation. In *BioNLP 2017*, pages 126–135, Vancouver, Canada,, August. Association for Computational Linguistics.

Linfeng Song, Daniel Gildea, Yue Zhang, Zhiguo Wang, and Jinsong Su. 2019. Semantic neural machine translation using amr. *Transactions of the Association for Computational Linguistics*, 7:19–31.

Pavlo Vasylenko, Pere Lluís Huguet Cabot, Abelardo Carlos Martínez Lorenzo, and Roberto Navigli. 2023. Incorporating graph information in transformer-based AMR parsing. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1995–2011, Toronto, Canada, July. Association for Computational Linguistics.

David Vilares and Carlos Gómez-Rodríguez. 2018. A transition-based algorithm for unrestricted AMR parsing. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 142–149, New Orleans, Louisiana, June. Association for Computational Linguistics.

Chuan Wang, Nianwen Xue, and Sameer Pradhan. 2015. A transition-based algorithm for AMR parsing. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 366–375, Denver, Colorado, May–June. Association for Computational Linguistics.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1212-1222, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Peiyi Wang, Liang Chen, Tianyu Liu, Damai Dai, Yunbo Cao, Baobao Chang, and Zhifang Sui. 2022. Hierarchical curriculum learning for AMR parsing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 333–339, Dublin, Ireland, May. Association for Computational Linguistics.

Chen Yu and Daniel Gildea. 2022. Sequence-to-sequence AMR parsing with ancestor information. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 571–577, Dublin, Ireland, May. Association for Computational Linguistics.

Zixuan Zhang and Heng Ji. 2021. Abstract Meaning Representation guided graph encoding and decoding for joint information extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 39–49, Online, June. Association for Computational Linguistics.

Sheng Zhang, Xutai Ma, Kevin Duh, and Benjamin Van Durme. 2019a. AMR parsing as sequence-to-graph transduction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 80–94, Florence, Italy, July. Association for Computational Linguistics.

Sheng Zhang, Xutai Ma, Kevin Duh, and Benjamin Van Durme. 2019b. Broad-coverage semantic parsing as transduction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3786–3798, Hong Kong, China, November. Association for Computational Linguistics.

Jiawei Zhou, Tahira Naseem, Ramón Fernandez Astudillo, and Radu Florian. 2021a. AMR parsing with action-pointer transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5585–5598, Online, June. Association for Computational Linguistics.

Jiawei Zhou, Tahira Naseem, Ramón Fernandez Astudillo, Young-Suk Lee, Radu Florian, and Salim Roukos. 2021b. Structure-aware fine-tuning of sequence-to-sequence transformers for transition-based AMR parsing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6279–6290, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

CCL 2024

# Automatic Construction of the English Sentence Pattern Structure Treebank for Chinese ESL learners

**Lin Zhu**[1,2*] **Meng Xu**[1,2*] **Wenya Guo**[1,2] **Jingsi Yu**[1,2]
**Liner Yang**[1,2†] **Zehuang Cao**[3] **Yuan Huang**[2] **Erhong Yang**[1,2]
[1]National Language Resources Monitoring and Research Center for Print Media,
Beijing Language and Culture University, China
[2]School of Information Science, Beijing Language and Culture University, China
[3]Faculty of Foreign Studies, Beijing Language and Culture University, China
lineryang@gmail.com

## Abstract

Analyzing long and complicated sentences has always been a priority and challenge in English learning. In order to conduct the parse of these sentences for Chinese English as Second Language (ESL) learners, we design the English Sentence Pattern Structure (ESPS) based on the Sentence Diagramming theory. Then, we automatically construct the English Sentence Pattern Structure Treebank (ESPST) through the method of rule conversion based on constituency structure and evaluate the conversion results. In addition, we set up two comparative experiments, using trained parser and large language models (LLMs). The results prove that the rule-based conversion approach is effective.

## 1 Introduction

Reading comprehension is a fundamental skill in English learning, pivotal for linguistic acquisition, critical thinking, and effective communication across various contexts. For Chinese ESL (English as a Second Language) learners, the ability to analyze complicated sentences represents both a central priority and a significant challenge in reading comprehension. To overcome this reading barrier, it is essential for learners to have a certain level of grammatical knowledge. Bernhardt (1993) believes that grammar is very crucial for second language learners' reading ability. Alderson(1993) considers grammatical ability an important foundation for second language learners' reading, emphasizing a vital to divide sentences into correct patterns.

Existing analysis tutorials for complicated English sentence typically contain only hundreds of example sentences, making it difficult for students to receive immediate and targeted feedback during practice, such as a book published by New Oriental Education, short for NOE300(Chen et al., 2019). Automatic syntactic analysis can compensate for this by transcending the boundaries of time and space and provide unlimited sentence analysis support. Most of the current automatic English grammar parses are designed for processing simple sentences, e.g. Enpuz[0] analyses sentences with an upper limit of 20 words in length. Based on these considerations, this paper conduct automatic grammar analysis of long and complicated sentences without length constraints. We adopts the widely recognized Sentence Diagramming theory, referring its more standardized approaches such as Grammar Revolution[1] and Sentence Analytics[2]. These methods have covered most English grammatical cases, providing vivid and detailed analysis, but their perspectives of grammar explanation are not completely suited to the learning habits of Chinese ESL learners. Therefore, we have made improvements such as grouping various clauses to-

[*]Equal contribution.
[†]Corresponding author: Liner Yang (mailto: lineryang@gmail.com)

©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

This work was supported by the MOE (Ministry of Education in China) Project of Humanities and Social Sciences (No. 23YJCZH264), the funds of Research Project of the National Language Commission (No. ZDA145-17) and Education and Teaching Reform Project of Beijing Language and Culture University in 2024 (No. JXGG202403).

[0]https://enpuz.com
[1]https://letsdiagram.com/
[2]https://www.sentenceanalytics.com/

gether, focusing on adverbial accompaniment, etc, and, designed the English Sentence Pattern Structure (ESPS).

Treebank is the processed corpus that records the syntactic annotation of every sentence, providing word segmentation, part-of-speech tagging, syntactic structure and other information. Based on the proposed grammatical system, we aim to build the English Sentence Pattern Structure Treebank (ESPST) and further develop automatic parses. The standard methods of constructing large-scale syntactic resources are manual annotation and automatic conversion. Manual annotation can ensure the data quality but is time-consuming and labor-intensive. A practical alternative method is to utilize existing treebank resources and converting them into the target treebank by finding the mapping relationships between two forms.

Current research on automatic treebank conversion mainly focuses on the conversion between constituency treebanks and dependency treebanks. Lin (1998) proposed an early method using a headword node table to convert constituency trees into dependency trees. Xia (2001) described two algorithms for converting constituency trees into dependency trees, employing a headword filtering table method, and proposed a new algorithm for converting the generated dependency trees back into constituency trees, with the results closely resembling the original Penn Treebank (PTB). Zabokrtsky (2003), Niu (2009), and Kong (2015) also conducted research on the conversion between constituency structure and dependency structure. In Chinese, some scholars have researched the conversion between constituency treebanks, dependency treebanks, and Chinese Sentence Pattern Structure Treebanks (SPST). Among them, Zhang (2018) converted the Tsinghua Chinese Treebank (TCT) into SPST, with an overall accuracy rate of 92.9%. Xie (2022) used rule-based methods to convert the Chinese Treebank (CTB) into SPST, with an overall accuracy rate of 89.72%. These studies have proved the feasibility of interconversion between different syntactic structures.

Using conversion rules and the advanced parser of the source treebank, we can automatically generate targeted trees from raw sentences. Yet, creating these conversion rules is a very challenging task that needs careful observation and steady practice, posing a significant challenge to our research. Considering the wide use of the PTB in natural language processing, we choose to convert the constituency structure treebank into ESPST. To verify the effectiveness of the conversion rules, we conducted experiments on a manually annotated test set to compare the conversion results with the effects of trained parser and large language models (LLMs). The results indicating that the rule-based conversion method proposed in this paper is the most effective.

## 2 Background

In this paper's conversion, the source treebank is mainly the English PTB corpus, and the target treebank is the ESPST we designed based on Sentence Diagramming theory, as introduced in Section 3. The formulation of conversion rules necessitates a comparative analysis of the grammatical forms between these two. This section will separately introduce PTB and Sentence Diagramming.

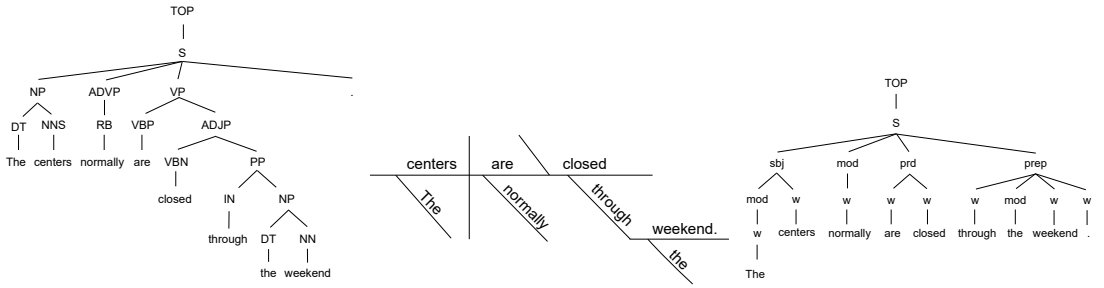
### 2.1 Penn Treebank

The English PTB corpus, particularly the section of the corpus corresponding to the articles of Wall Street Journal (WSJ), is one of the most known and used constituency structure corpus for the evaluation of models for sequence labelling. It is a corpus consisting of over 4.5 million words of American English. The material annotated by PTB includes such wide-ranging genres as IBM computer manuals, nursing notes, Wall Street Journal articles, and transcribed telephone conversations. The large amount of data produced by the project continues to provide an available resource for computational linguists, natural language programmers, corpus linguists and others interested in empirical language studies.

According to Marcus(1993), PTB corpus is annotated for part-of-speech (POS) information and skeletal syntactic structure. Considering the notable differences between the syntactic structure annotations provided in the PTB and the grammar accustomed by Chinese ESL learners, this paper exclusively focuses on the POS annotations of the PTB. The majority of the output of the PTB consists of tagged and bracketed versions. As shown in Figure 1a, compared to the ESPST as proposed in Section 3, the PTB

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1223-1238, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

we focus on only annotates part-of-speech information and hierarchical structure information, lacking in the depiction of syntactic relationships between sentence components, which can be extracted based on rules.




(a) Constituency Structure (b) Sentence Diagramming (c) English Sentence Pattern Structure

Figure 1: Example of three formats.

### 2.2 Sentence Diagramming

Sentence Diagramming is a pictorical representation of a sentence's grammatical structure, which is used to teaching difficult written language. The model shows the relations between words and the nature of sentence structure and can be used to help recognize which potential sentences are grammatically correct.

Most Sentence Diagramming methods in pedagogy are based on the work of Alonza Reed and Brainerd Kellog(1886). Sentences in the Reed-Kellogg system are diagrammed according to the following forms: the diagram begins with a horizontal line called the base; the subject is written on the left, the predicate on the right, separated by a vertical bar; the verb and its object are separated by a line that ends at the baseline; modifiers, as well as prepositional phrases, are placed on slanted lines below the word they modify. These basic diagramming conventions are augmented for other types of sentence structures, e.g. for coordination and subordinate clauses.

A specific example of sentence diagramming is illustrated in Figure 1b. Based on the direct modifying function of adverbials on the predicate meaning, *normally* is attached below the predicate *are*. Above the horizontal line is the simplified main component of the sentence, "centers are closed".

To further deepen our understanding of the components and rules of Sentence Diagramming, we referenced an exceptional work. The Grammar Revolution project, developed by Elizabeth O'Brien, aims to redefine traditional methods of grammar learning by offering an innovative perspective through interest. In Grammar Revolution, 11 lessons unfold sequentially: Basic Sentence Diagramming, Modifiers, Prepositional Phrases, Coordinating Conjunctions, etc. By transforming abstract grammatical concepts into concrete, visual patterns, this project not only makes grammar learning more engaging but also opens new avenues for learners who feel intimidated by or disinterested before. This has bolstered our confidence in applying Sentence Diagramming theory to the research for Chinese ESL learners.

## 3 The English Sentence Pattern Structure

Based on the theories of Sentence Diagramming discussed in Section 2.2, and in conjunction with Chinese ESL learners' cognitive habits, we have defined 14 grammatical labels involving sentence components and logical relationships. As shown in the Table 1, the 14 labels are categorized as main components, supplementary components, and relational components. These 14 components can all be represented by diagrams, which will not be elaborated here.

### 3.1 Main Components

For every sentence or clause, we analyze its main components, including the subject, predicate, direct object, indirect object, and predicative. These five main components constitute three basic sentence patterns: subject-verb-object, subject-verb-indirect object-direct object, and subject-linking verb-predicative.

| Categories | Labels | Explanation | Categories | Labels | Explanation |
|---|---|---|---|---|---|
| Main Components | sbj | Subject | Supplementary Components | mod | Modification |
| | prd | Predicate | | advcla | Adverbial Accompaniment |
| | obj | Object | | cla | Clause |
| | pred | Predicative | | wh | Relative Connectives |
| | iobj | Indirect Object | Relational Components | sencoo | Sentential Coordination |
| Supplementary Components | todo | Todo Infinitive | | phrcoo | Phrasal Coordination |
| | prep | Prepositional Phrase | | cc | Coordinating Conjunction |

Table 1: The 14 grammatical labels.

The predicate refers to the main verb and its modifying elements. For example, in the sentence *He has not yet seen the bird*, the predicate would be *has not yet seen*, which includes the verb *has* and *seen* carrying tense information and modifiers *not yet*.

The linking verbs that introduce predicative indicates the subject's state, quality, characteristics, or nature, including verbs such as *be, remain, feel*, and their variant forms. Moreover, the predicative may be a noun, adjective, certain adverbs, non-finite verbs, prepositional phrases, or clauses.

### 3.2 Supplementary Components

To facilitate learners' grasp of the main skeleton of sentences, the *mod* label encompasses various components with a modifying function without further detailed subdivision, which includes adverbs, adjectives, numerals, quantity phrases, possessive pronouns, post-modifiers led by gerunds or past participles, appositives, etc.

Infinitives are typically used to express purpose or intention or as a complement to another verb. Prepositional phrases act as adverbials of time, place, manner, etc. For example, in *I came here to see the exhibition*, the *to* leads an infinitive, indicating the purpose of coming here is to see the exhibition; in *I look forward to seeing you soon*, *to* is a preposition as part of the phrase *look forward to*, followed by the gerund form *seeing*, rather than an infinitive.

Adverbial accompaniment, describing subsidiary actions or states that occur concurrently with the main action, is an integral part in sentence. It can be expressed through various grammatical forms, such as present participle phrases and past participle phrases. Formally, when adverbial clauses are positioned at the beginning or middle of a sentence, they are often separated from the main clause by commas.

Various types of clauses, such as adverb clauses, adjective clauses, noun clauses, etc., are uniformly classified as *cla*. The relative connectives of clauses can be a single word, such as *because, if, when, although*, or phrases such as *even though, in order that*.

### 3.3 Relational Components

In the ESPS, we further define the logical relationships of *coordination*. Coordination refers to the structural equivalence of two or more sentence components, which jointly function as a more significant unit and semantically represent various meanings such as *alliance*, *contrast*, and *progression*. Within a sentence, the coordination of sub-sentences or clauses is defined as *sencoo*, and the coordination of phrases is defined as *phrcoo*, with the coordinating conjunctions guiding these two types of relationships defined as *cc* (such as *and* or *but*). To enhance parsing efficiency, for phrase coordination, we currently focus only on the coordination of subjects, predicates, predicatives, and objects, which are directly related to the basic structure and meaning.

## 4 Constructing the English Sentence Pattern Structure Treebank

In English, the techniques of constituency structure are relatively mature and have yielded promising results. Therefore, we choose the constituency structure as the source treebank and construct the ESPST through rule-based conversion. Specifically, this involves formulating conversion rules for the 14

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1223-1238, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

| Labels | Explanation | constituency Structure | Conversion Rules | Examples |
|---|---|---|---|---|
| sbj | subject | S → NP | <sbj>NP</sbj> | It wasn't Black Monday. |
| prd | predicate | S/SINV → VP → (VP → VERB) | <prd>VERB</prd> | The equity market was illiquid. |
| obj | object | S/SINV → VP → (VP → VERB + NP) | <obj>NP</obj> | They received approvals for development. |
| pred | predicative | VP → (VERB + NP/PP/ ADVP/ADJP/SBAR) and VERB is linking verb | <pred>NP/PP/ADVP/ ADJP/SBAR</pred> | It wasn't Black Monday. |
| iobj | indirect object | S/SINV → VP → (VP → VERB + NP1 + NP2) | <iobj>NP1</iobj> <obj>NP2</obj> | She gave me a book. |

Table 2: Sample conversion rules of main components. VERB includes the labels of *VBP, MD, VBD, VBZ, VBN, VB,* and *VBG*.

grammatical labels described in Section 3 and handling certain exceptional cases. However, compared to ESPST, the PTB lacks the depiction of syntactic relationships between sentence components. Thus, formulating the rules can also be considered as the precise correspondence between part-of-speech information and syntactic component information in English. The following elaborates on the conversion rules we have developed.

### 4.1 Conversion Rules of Grammatical Components

In the constituency structure, information on sentence components is scattered among part-of-speech labels, which cannot be directly correlated on a one-to-one basis. Based on this, we have compiled detailed rules for converting grammatical components. For each label, we only present a representative conversion rule here, with other specific rules available in the appendix A.

#### 4.1.1 Conversion of Main Components

The selected rules for converting the five main components from constituency structure is shown in Table 2.

In the constituency structure, the noun phrase *NP* under the sentence *S* and the inverted sentence *SINV* is typically the subject of the sentence. If there is no *NP* at this position, then a sentence *S* or a clause *SBAR* at the same level is matched as the subject.

Three rules for predicate conversion correspond to scenarios where the same predicate part has one, two, or three verbs. These scenarios involve different hierarchical relationships in the constituency structure tree.

The conversion rules for direct objects, indirect objects, and predicatives are closely related to the rules for predicate: within the same level after a predicate in the constituency structure, if there is one *NP*, it is matched as a single direct object; if there are two *NP*, the first is matched as an indirect object and the latter as a direct object; *NP, PP, ADVP, ADJP, SBAR* at the same level as the linking verb are matched as predicatives.

#### 4.1.2 Conversion of Supplementary Components

The selected rules for converting the six supplementary components from constituency structure is shown in Table 3.

For the preposition *to*, if its parent node is a verb phrase (*VP*), then this verb phrase matches as *todo*; if its parent node is a prepositional phrase (*PP*), then this prepositional phrase matches as *prep*. Moreover, combinations of other prepositions with noun phrases, sentences, or adjective phrases also match as *prep*.

As previously mentioned, the *mod* label encompasses various components with a modifying function. The conversion rules correspond to these nine types of modifiers: adverbs and phrases serving as adverbials, adjectives and adjective phrases, numerals, and quantity phrases, possessive pronouns, nouns

CCL 2024

| Labels | Explanation | constituency Structure | Conversion Rules | Examples |
|---|---|---|---|---|
| todo | Todo Infinitive | S/SBAR → VP → TO + VP | <todo>VP</todo> | And the link with stocks began to fray again. |
| prep | Prepositional Phrase | PP → IN/TO + NP/S/ADJP | <prep>PP</prep> | At the end of the day, 251.2 million shares were traded. |
| mod | Modification | NP/NML → JJ/JJS/ ADJP/RBR/PDT + NP/NN/NNS/NNP/NNPS | <mod>JJ/JJS/ADJP/ RBR/PDT</mod> | I wouldn't expect an immediate resolution to anything. |
| advcla | Adverbial of Accompaniment | S1/VP1 → S2/VP2 → PP/VP3 → VBG/VBN + XP | <advcla>S2/VP2</advcla> | Noting others' estimates, he said October. |
| cla | Clause | SBAR | <cla>SBAR</cla> | When the dollar is in a free-fall, even central banks can't stop it. |
| wh | Relative Connectives | SBAR → WHNP | <wh>WHNP</wh> | Speculators are calling for a degree of liquidity that is not there in the market. |

Table 3: Sample conversion rules of supplementary components. *XP* stands for any component.

| Labels | Explanation | constituency Structure | Conversion Rules | Examples |
|---|---|---|---|---|
| sencoo | Sentential Coordination | S1 → S2 + CC + S3 | <sencoo>S1</sencoo> | But the build-up of S&P futures sell orders weighed on the market, and the link withstocks began to fray again. |
| phrcoo | Phrasal Coordination | S → NP1 → NP2 + CC + NP3 | <phrcoo>NP1</phrcoo> | Many money managers and some traders had already left their offices. |
| cc | Coordinating Conjunction | CC in sencoo/phrcoo | <cc>CC</cc> | Many money managers and some traders had already left their offices. |

Table 4: Sample conversion rules of relational components.

or noun phrases modifying another noun, post-modifiers led by gerunds or past participles, appositives, reflexive pronouns.

Adverbial accompaniments are typically led by the present participle *VBG* or past participle *VBN*. Comsidering the characteristics of their parent nodes, we define the conversion rule as *S/VP → S1/VP1 → PP/VP2 → VBG/VBN + XP*.

The *cla* label directly corresponds to the *SBAR* label in the PTB. The *wh* label appear in various forms, which can be single words, such as *IN* in *SBAR → IN + S*; or phrases, such as *WHNP* in *SBAR → WHNP*.

#### 4.1.3 Conversion of Relational Components

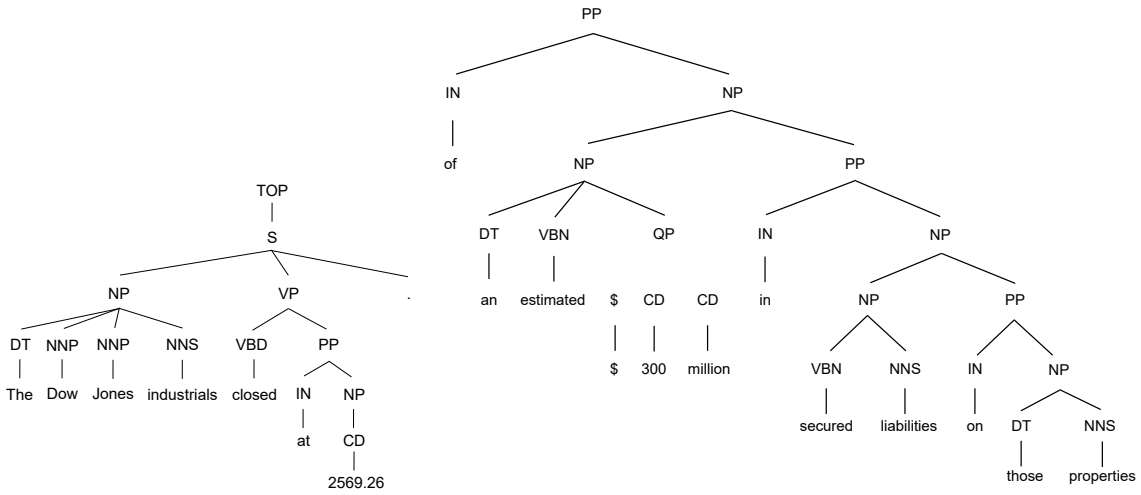
The selected rules for converting the three relational components from constituency structure is shown in Table 4.

Sentence coordination encompasses coordination of sub-sentence and clauses, while phrase coordination encompasses coordination of predicates, subjects, objects, and predicatives. The conjunctions of sentence coordination and phrase coordination are matched as coordinating conjunctions.

### 4.2 Conversion Rules of Special Cases

In practice, we found that although the conversion rules cover the majority of standard structures, there are several special cases require appropriate handling.

- When the headword (i.e., the main noun in a noun phrase) has multiple modifiers, these modifiers should all point to the last noun individually. Noun phrases often follow a certain hierarchical structure, where the noun placed last (except in some cases of post-modifying attributes) is the




(a) Special Case of Proper Nouns.

(b) Special Case of Nest Restrictions.

Figure 2: Examples of special cases.

| Dataset | Source | Numbers | Average length | Train | Dev | Test |
|---|---|---|---|---|---|---|
| PTB | Wall Street Journal articles, etc. | 800 | 24.43 | 640 | 80 | 80 |
| NOE | GRE, GMAT, LSAT, etc. | 200 | 41.26 | 160 | 20 | 20 |
| Total | - | 1000 | 27.79 | 800 | 100 | 100 |

Table 5: Statistics of datasets.

headword, with preceding modifiers sequentially modifying and specifying it. Adding this rule aids in maintaining the grammatical correctness of the sentence.

- We stipulate that adjacent *NNP/NNPS* (proper nouns in PTB) are considered a joint unit for division or matching. In sentences containing place names, personal names, or specific terms, when two or more proper nouns are closely connected, they usually form a single semantic unit, expressing a compound concept or a concrete entity. For instance, as shown in Figure 2a, the modifier of the noun *industrials* includes *The* and *Dow Jones*, rather than *The*, *Dow*, and *Jones* as three separate modifiers.

- To maintain the clarity of sentence structure, we have imposed restrictions on the nesting of certain labels. The specific rules are as follows: *prep* and *todo* do not nest within *prep* or *todo* but can nest other labels; *wh* do not nest any other labels. For instance, in Figure 2b, *of an estimated $300 million in secured liabilities on those properties* would be converted to a *prep* label, without further conversion for *in secured liabilities on those properties* and *on those properties*. This approach is adopted to prevent the potential for ambiguity, which can arise from complex tag nesting structures in handling complicated sentences.

## 5 Experiments

Based on the rules, we completed the conversion of ESPST for a total of 39,406 sentences in constituency structure and manually annotated 1000 sentences as a test set to evaluate the conversion results. Among the 1000 sentences, we calculated the annotation consistency to measure the rules and data quality. To further verify the effectiveness of the treebank conversion method, we set up two additional experiments: a trained parser and a LLMs analysis.

CCL 2024

| Categories | Labels | P | R | F1 | Categories | Labels | P | R | F1 |
|---|---|---|---|---|---|---|---|---|---|
| Main Components | sbj | 99.22 | 97.10 | 98.15 | Supplementary Components | todo | 98.06 | 95.60 | 96.82 |
| | prd | 92.53 | 88.86 | 90.66 | | advcla | 9.82 | 73.49 | 17.33 |
| | pred | 58.98 | 64.17 | 61.47 | | cla | 96.77 | 97.82 | 97.29 |
| | obj | 86.19 | 90.71 | 88.39 | | wh | 97.69 | 96.12 | 96.90 |
| | iobj | 38.64 | 85.00 | 53.12 | Relational Components | sencoo | 98.72 | 92.77 | 95.65 |
| Supplementary Components | mod | 86.87 | 84.77 | 85.81 | | phrcoo | 99.30 | 78.33 | 87.58 |
| | prep | 95.16 | 94.36 | 94.76 | | cc | 96.82 | 82.56 | 89.12 |
| Avg. P | | | | | 82.48 | | | | |
| Avg. R | | | | | 87.26 | | | | |
| Avg. F1 | | | | | 84.81 | | | | |

Table 6: Main results of rule-based conversion in 1000 sentences.

### 5.1 Dataset

The total 39,406 sentences in constituency structure are from the test division of PTB. Our dataset for evaluating the conversion results consists of 1000 sentences including 800 from PTB23 and 200 from NOE300. NOE300 is a book compiled to help Chinese students with reading long and complex sentences in tests, selecting examples of such sentences that appear in the GRE (Graduate Record Examinations), GMAT (Graduate Management Admission Test), and LSAT (Law School Admission Test). For the subsequent two comparative experiments in Section 5.3, we constructed the train, validation, and test sets from the dataset in an 8:1:1 ratio. The specific information about the 1000 dataset is shown in Table 5.

### 5.2 Evaluation of Rule Conversion

We manually annotated the ESPS of the 1000 sentences and calculated the Fleiss' Kappa score (Fleiss, 1971) of annotation agreement on two annotators to be 0.88, indicating that the grammatical labels are scientifically sound. By categorizing the components, we examine the specific results of automatic conversion, with overall result presented in the Table 6 below and detailed data available in the appendix B.

Our findings include:

1. The overall conversion results are satisfactory, indicating that the conversion rules for handling the sentence's main components, supplementary components, and relational components are scientifically valid, resulting in a high-quality treebank. The F1 scores on both datasets exceed 80, demonstrating the transfer-ability of this method to texts in other domains.

2. Conversion results for PTB23 data outperform those for NOE300 data. Under the same conversion rules, the F1 score for 800 PTB23 data conversions is 86.61, while for 200 NOE300 data conversions is 80.82, a difference of 5.79. As for the reason, the latter's sentences are, on average, about twice as long as the former's and are grammatically more complex, potentially leading to cases not covered by certain rules. Moreover, the constituency structure form of NOE300 used to generate conversion results was produced by the Berkeley Constituency Parser(Kitaev and Klein, 2018), which may introduce bias.

3. Among the 14 components, subjects, predicate, prepositional phrases, infinitives, clause, sentence coordination, and relative connectives have the better conversion results, with overall F1 scores exceeding 90. This indicates that their conversion rules can cover most grammatical cases, with constituency structure corresponding accurately to the respective ESPS, resulting in a low error rate in conversion results. Additionally, conversion results for components like object, modification, phrasal coordination, and coordinating conjunction are also considerable.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1223-1238, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

4. Predicatives in the PTB treebank appear in various forms, including noun phrases, adjective phrases, adverbial phrases, and prepositional phrases, for which we have written conversion rules to match these cases. However, in many sentences' constituency structures, nodes at the same level as and following the linking verb can often match multiple predicatives, only one of which is correct. For example, in the sentence *Diamond Shamrock is the operator, with a 100% interest in the well*, conversion rules match two predicatives of noun phrases and prepositional phrases, but only the former is correct. This issue may explain the relatively worse conversion results for predicatives.

5. In constituency structure trees, many temporal adverbials appear as NP(noun phrases), affecting conversion results for components like direct and indirect objects. In particular, indirect objects are significantly impacted, with high recall but low precision due to the small base. Observations reveal many temporal adverbials such as *last week*, and*yesterday*, *tomorrow morning* being matched as indirect objects. This issue can be partially resolved by defining a list of prohibited words: among the 155 misclassifications, 59 are temporal adverbials, indicating that introducing a prohibited list of temporal adverbials could resolve about one-third of this kind of issues.

6. Concomitant adverbials are form-flexible, making the conversion task challenging. Conversion rules matching verb phrases led by present or past participles yield many wrong components, such as predicates. In 1000 sentences, the label appeared 61 times in the manually annotated results but over 200 times in the conversion results, indicating a lower accuracy rate in the conversion. This suggests the weak correspondence between such sentence structure labels and part-of-speech information requires alternative matching approaches.

### 5.3 Comparative Experiments

We conducted two sets of experiments to compare the effectiveness of the treebank conversion rules proposed in this paper.

#### 5.3.1 Setup

**Experiment 1**: This experiment primarily investigates the performance of an automatic syntactic parser trained on ESPST generated through rule-based conversion, aiming to explore the practical value of the method we propose. Drawing on Kitaev's neural network model[3](Kitaev et al., 2018) based on self-attention mechanisms, we trained an automatic syntactic parser for ESPS. The training set is the ESPST of 39,406 sentences converted from constituency structure trees in PTB, and the testing set is the 1000 sentences introduced in Section 5.1. The model employs an encoder-decoder architecture, using the pre-trained model Bert for the encoding phase and incorporating part-of-speech and positional information as auxiliary inputs to the model. The encoder sums the word representations $[w_1, ..., w_n]$, part-of-speech representations $[m_1, ..., m_n]$, and positional representations $[p_1, ..., p_n]$ to obtain word embeddings, which are then encoded using a multi-head attention mechanism. The decoder employs the CKY algorithm (Kasami, 1966; Younger, 1967; Cocke, 1969) to generate the ESPST.

**Experiment 2**: To test the syntactic analysis capability of LLMs on complex English sentences, we conducted prompt-based experiments on the general-domain GPT-4. The testing set is also the 1000 sentences introduced in Section 5.1. The full prompt (shown in Appendix C) given to GPT-4 for each testing sentence consisted of the following ordered elements:

- **Syntactic Labels**, introducing the 14 grammar labels and some necessary explanations;
- **Special Rules**, describing the special rules, which significantly impact the generation results;
- **Task**, explaining the task of analyzing sentences based on the ESPS;
- **Examples**, giving three complex English sentences and their correct output results. These three sentences are not included in the test set. These examples provided the LLMs with detailed information on the output format and the handling of punctuation;

[3]https://github.com/nikitakit/self-attentive-parser

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1223-1238, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

| Models | PTB | | | NOE | | | TOTAL | | |
|---|---|---|---|---|---|---|---|---|---|
| | P | R | F1 | P | R | F1 | P | R | F1 |
| Trained Parser | 85.21 | 85.35 | 85.28 | 79.73 | 78.37 | 79.04 | 82.92 | 82.69 | 82.80 |
| GPT-4 | 41.38 | 38.51 | 39.89 | 35.31 | 33.45 | 34.35 | 40.17 | 37.50 | 38.78 |
| Rule-based Conversion | 83.44 | 90.04 | 86.61 | 80.55 | 81.10 | 80.82 | 82.48 | 87.26 | 84.81 |

Table 7: Results of comparative experiments in 1000 test dataset.

- **Testing Data**, giving the 1000 sentences to be tested, one at a time.

For the 1000 responses returned by GPT-4, we first extract the parenthetical syntactic trees to clean the data. Upon observation, the pairing of parentheses in these responses are not standardized, where there are cases of missing or redundant parentheses. Therefore, we resort to manual proofreading to adjust the format before calculate the results.

### 5.3.2 Results and Analysis

The overall performance of methods of trained parser, LLMs analysis, and rule-based conversion are listed in Table 7.

The conclusions drawn from the table are as follows: The method proposed in this paper, rule-based conversion from constituency structure, shows the best effect, with an F1 value 2.01 higher than that of the trained parser and 46.63 higher than the LLMs analysis. This indicating that the rule-based conversion algorithm has certain advantages in automatic treebank construction, and data generated through conversion rules demonstrates significant utility in training parses.

Compared to the trained parser, the method based on conversion rules does not rely on the manually annotated 1000 sentences, allowing it to be transferred to texts in other domains, showing more robust universality. The LLMs analysis experiment performed poorly, possibly due to the general-domain LLMs' lower accuracy in complex sentence syntactic analysis tasks or the model's unfamiliarity with the grammar labeling system tailored for Chinese ESL learners. Additionally, the length of the test sentences and the bracket format of the treebank may have contributed to the reduced accuracy.

In summary, the rule-based conversion algorithm proposed in this paper has certain advantages in the automatic construction of ESPST, showing vital accuracy and universality in analysis.

## 6 Conclusion

We developed an ESPS rooted in Sentence Diagramming theory, which is suited to Chinese ESL learners. Through rule-based conversion from the PTB, we constructed the ESPST and evaluated its effectiveness. Comparative experiments, including parser training and LLMs analysis, showed that our treebank conversion rule-based method yielded the best results. This work provides a new perspective on efficient English grammar learning of long and complicated sentence. However, our conversion process have shortcomings, such as the rules for indirect objects, predicatives, and adverbial clauses, which need further refinement. In the future, we will continue to optimize the conversion results and build an analysis platform for ESPS, realizing the visualization of automatic syntactic analysis.

CCL 2024

## References

Elizabeth B Bernhardt. 1993. *Reading development in a second language: Theoretical, empirical, & classroom perspectives.* ERIC.

Qi Chen, Yi Ge, and Yuzhen Yan. 2019. *New Oriental Education: 300 Detailed Explanations and Practices for Long and Complex GRE/GMAT/LSAT Sentences*. Zhejiang Education Publishing House.

John Cocke. 1969. *Programming languages and their compilers: Preliminary notes*. New York University.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Tadao Kasami. 1966. An efficient recognition and syntax-analysis algorithm for context-free languages. *Coordinated Science Laboratory Report no. R-257*.

Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. *arXiv preprint arXiv:1805.01052*.

Nikita Kitaev, Steven Cao, and Dan Klein. 2018. Multilingual constituency parsing with self-attention and pre-training. *arXiv preprint arXiv:1812.11760*.

Lingpeng Kong, Alexander M Rush, and Noah A Smith. 2015. Transforming dependencies into phrase structures. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 788–798.

Dekang Lin. 1998. A dependency-based method for evaluating broad-coverage parsers. *Natural Language Engineering*, 4(2):97–114.

Mitch Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330.

Zheng-Yu Niu, Haifeng Wang, and Hua Wu. 2009. Exploiting heterogeneous treebanks for parsing. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 46–54.

Alonzo Reed and Brainerd Kellogg. 1886. *Higher lessons in English*. Scholars' Facsimiles & Reprints.

Dianne Wall and J Charles Alderson. 1993. Examining washback: the sri lankan impact study. *Language testing*, 10(1):41–69.

Fei Xia and Martha Palmer. 2001. Converting dependency structures to phrase structures. In *Proceedings of the first international conference on Human language technology research*.

Chenhui Xie, Zhengsheng Hu, Liner Yang, Tianxin Liao, and Erhong Yang. 2022. Automatic construction of sentence pattern structure treebank. In *Proceedings of the 21st Chinese National Conference on Computational Linguistics*, pages 464–474.

Daniel H Younger. 1967. Recognition and parsing of context-free languages in time n3. *Information and control*, 10(2):189–208.

Zdeněk Žabokrtskỳ and Otakar Smrz. 2003. Arabic syntactic trees: from constituency to dependency. In *10th Conference of the European Chapter of the Association for Computational Linguistics*.

YinBing Zhang, Jihua Song, Weiming Peng, Yawei Zhao, and Tianbao Song. 2018. Automatic conversion of phrase structure treebank to sentence structure treebank. *Journal of Chinese Information Processing*, (5):31–41.

CCL 2024

## A Full Conversion Rules

Full conversion rules in this work is shown in Table 8-10.

| Labels | Explanation | constituency Structure | Conversion Rules | Examples |
|---|---|---|---|---|
| sbj | subject | S → NP | <sbj>NP</sbj> | It wasn't Black Monday. |
| | | SINV → NP | <sbj>NP</sbj> | At 2:43 p.m. EDT, came the sickening news. |
| | | S → S/SBAR + VP | <sbj>S/SBAR</sbj> | The thing that they have done is a question. |
| prd | predicate | S/SINV → VP → (VP → VERB) | <prd>VERB</prd> | The equity market was illiquid. |
| | | S/SINV → VP → (VP → VERB1 + VP→ (VP → VERB2)) | <prd>VERB1 + VERB2</prd> | At the end of the day, 251.2 million shares were traded. |
| | | S/SINV → VP → (VP → VERB1 + VP → (VP → VERB2 + VP → (VP → VERB3))) | <prd>VERB1 + VERB2 + VERB3</prd> | Several traders could be seen shaking their heads. |
| obj | object | S/SINV → VP → (VP → VERB + NP) | <obj>NP</obj> | They received approvals for development. |
| | | S/SINV → VP → (VP → VERB + VP → (VP → VERB + NP)) | <obj>NP</obj> | He could watch updates on prices and pending stock orders. |
| | | S/SINV → VP → (VP → VERB + VP → (VP → VERB + VP → (VP → VERB + NP))) | <obj>NP</obj> | The suppliers haven't been filling their quotas to the full extent. |
| pred | predicative | VP → (VERB + NP/PP/ ADVP/ADJP/SBAR) and VERB is linking verb | <pred>NP/PP/ADVP/ ADJP/SBAR</pred> | It wasn't Black Monday. |
| iobj | indirect object | S/SINV → VP → (VP → VERB + NP1 + NP2) | <iobj>NP1</iobj> <obj>NP2</obj> | She gave me a book. |
| | | S/SINV → VP → (VP → VERB + VP → (VP → VERB + NP1 + NP2)) | <iobj>NP1</iobj> <obj>NP2</obj> | She has told us the news. |
| | | S/SINV → VP → (VP → VERB + VP → (VP → VERB + VP → (VP → VERB + NP1 + NP2))) | <iobj>NP1</iobj> <obj>NP2</obj> | He might have offered his colleague some help. |

Table 8: Full conversion rules of main components. VERB includes the labels of *VBP, MD, VBD, VBZ, VBN, VB,* and *VBG*.

CCL 2024

| Labels | Explanation | constituency Structure | Conversion Rules | Examples |
|---|---|---|---|---|
| todo | Todo Infinitive | S/SBAR → VP → TO + VP | <todo>VP</todo> | And the link with stocks began to fray again. |
| prep | Prepositional Phrase | PP → IN/TO + NP/S/ADJP | <prep>PP</prep> | At the end of the day, 251.2 million shares were traded. |
| mod | Modification | NP/NML → JJ/JJS/ ADJP/RBR/PDT + NN/NNS/NP/NNP/NNPS | <mod>JJ/JJS/ADJP/ RBR/PDT</mod> | I wouldn't expect an immediate resolution to anything. |
| | | (NP/ADJP/PP/S/INTJ/ VP → RB)/ADVP | <mod>RB/ADVP</mod> | These stocks eventually reopened. |
| | | NP → DT + XP | <mod>DT</mod> | The equity market was illiquid. |
| | | NP → NN1/NP1/NNP1/ VBP/VBG/NNS1 + NN2/NNS2/NP2/NNP2 | <mod>NN1/NP1/NNP/ VBP/VBG/NNS1</mod> | The equity market was illiquid. |
| | | NP → NP1+ ,/: + NP2 + punctuation | <mod>NP2</mod> | He is Howard Rubel, an analyst at Lawrance Inc. . |
| | | NP → QP/CD + NNS/NN | <mod>QP/CD</mod> | At the end of the day, 251.2 million shares were traded. |
| | | NP→ PRP$ + NN/NNP/ NNS/NP/NNPS/VBG | <mod>PRP$</mod> | Several traders could be seen shaking their heads. |
| | | NP → NP1/PP → VBG/VBN + XP | <mod>NP1/PP</mod> | The book lying on the table is mine. |
| | | NP → NP1 + NP2 and Reflexive pronouns in NP2 | <mod>NP2</mod> | It is index of the stock market itself. |
| advcla | Adverbial of Accompaniment | S1/VP1 → S2/VP2 → PP/VP3 → VBG/VBN + XP | <advcla>S2/VP2</advcla> | Noting others' estimates, he said October. |
| cla | Clause | SBAR | <cla>SBAR</cla> | When the dollar is in a free-fall, even central banks can't stop it. |
| wh | Relative Connectives | SBAR → WHNP | <wh>WHNP</wh> | Speculators are calling for a degree of liquidity that is not there in the market. |
| | | SBAR → IN + S | <wh>IN</wh> | There came news that the UAL group couldn't get financing for its bid. |
| | | SBAR → WHADVP | <wh>WHADVP</wh> | When the dollar is in a free-fall, even central banks can't stop it. |
| | | SBAR → WHPP | <wh>WHPP</wh> | But nobody knows at what levell the futures and stocks will open today. |

Table 9: Full conversion rules of supplementary components. *XP* stands for any component.

CCL 2024

| Labels | Explanation | constituency Structure | Conversion Rules | Examples |
|---|---|---|---|---|
| sencoo | Sentential Coordination | S1 → S2 + CC + S3 | <sencoo>S1</sencoo> | But the build-up of S&P futures sell orders weighed on the market, and the link withstocks began to fray again. |
| | | SBAR1 → SBAR2 + CC + SBAR3 | <sencoo>SBAR1</sencoo> | He said that he had not yet seen the bid but that he would review it. |
| phrcoo | Phrasal Coordination | S → NP1 → NP2 + CC + NP3 | <phrcoo>NP1</phrcoo> | Many money managers and some traders had already left their offices. |
| | | S → VP → (VP + CC + VP) / (VBD + CC + VBD) | <phrcoo>VP</phrcoo> | Mr. Shidler's company specializes in commercial real-estate investment and claims to have $1billion in assets. |
| | | VP → VERB + (NP1 → NP2 + CC + NP3) | <phrcoo>NP1</phrcoo> | A portion will be used to repay its bank debt and other obligations. |
| | | VP → VERB + (NP1 → NP2 + CC + NP3)/ (PP1 → PP2 + CC + PP3)/ (ADJP1 → ADJP2/JJ1 + CC + ADJP3/JJ2)/ (ADVP1 → ADVP2/RB1 + CC + ADVP3/RB2) and VERB is linking verb | <phrcoo>NP1/PP1/ ADJP1/ADVP1</phrcoo> | Mr.Simpson is a developer and a former senior executive of LJ. Hooker. |
| cc | Coordinating Conjunction | CC in sencoo/phrcoo | <cc>CC</cc> | Many money managers and some traders had already left their offices. |

Table 10: Full conversion rules of relational components.

CCL 2024

## B Detailed Conversion Results

The specific performance of the conversion rules on PTB23 and NOE300 is shown in Table 11.

| | Categories | Labels | P | R | F1 | Categories | Labels | P | R | F1 |
|---|---|---|---|---|---|---|---|---|---|---|
| PTB23 | Main Components | sbj | 99.24 | 98.02 | 98.62 | Supplementary Components | todo | 100.00 | 100.00 | 100.00 |
| | | prd | 92.81 | 89.90 | 91.33 | | advcla | 10.94 | 73.53 | 19.05 |
| | | pred | 60.58 | 65.97 | 63.16 | | cla | 99.84 | 99.69 | 99.77 |
| | | obj | 84.46 | 91.20 | 87.70 | | wh | 99.75 | 98.27 | 99.00 |
| | | iobj | 38.89 | 100.00 | 56.00 | Relational Components | sencoo | 100.00 | 94.34 | 97.09 |
| | Supplementary Components | mod | 86.44 | 85.31 | 85.87 | | phrcoo | 99.03 | 82.26 | 89.87 |
| | | prep | 99.39 | 96.56 | 97.96 | | cc | 96.73 | 85.55 | 90.80 |
| | Avg. P | | | | | 83.44 | | | | |
| | Avg. R | | | | | 90.04 | | | | |
| | Avg. F1 | | | | | 86.61 | | | | |
| | Categories | Labels | P | R | F1 | Categories | Labels | P | R | F1 |
| NOE300 | Main Components | sbj | 99.17 | 94.46 | 96.75 | Supplementary Components | todo | 92.31 | 83.72 | 87.80 |
| | | prd | 91.81 | 86.25 | 88.95 | | advcla | 6.17 | 73.33 | 12.29 |
| | | pred | 56.35 | 61.21 | 58.68 | | cla | 91.56 | 94.55 | 93.03 |
| | | obj | 90.43 | 89.60 | 90.02 | | wh | 95.27 | 93.60 | 94.43 |
| | | iobj | 37.50 | 50.00 | 42.86 | Relational Components | sencoo | 96.43 | 90.00 | 93.10 |
| | Supplementary Components | mod | 87.82 | 83.62 | 85.67 | | phrcoo | 100.00 | 69.64 | 82.11 |
| | | prep | 85.28 | 88.87 | 87.04 | | cc | 97.01 | 76.47 | 85.53 |
| | Avg. P | | | | | 80.55 | | | | |
| | Avg. R | | | | | 81.10 | | | | |
| | Avg. F1 | | | | | 80.82 | | | | |

Table 11: Detailed conversion results.

CCL 2024

## C Prompt in LLMs Experiment

The full prompt in the LLMs experiment is shown in Figure 3.

**Prompt**

Syntactic Labels: Subject <sbj>, Predicate <prd>, Object <obj>, Indirect Object <iobj>, Predicative <pred>, Infinitive <todo>, Prepositional Phrase <prep>, Clause <cla>, Modifier <mod> (adverbs and phrases as adverbials, adjectives, adjective phrases, numerals and quantifier phrases, possessive pronouns, nouns or noun phrases modifying another noun, post-modifiers led by gerunds or past participles, appositives, reflexive pronouns, etc.), Subordinating Conjunction <wh>, Coordinating Conjunction <cc>, Sentence Coordination <sencoo>, Phrase Coordination <phrcoo> (coordination of subjects, predicates, predicatives, and objects), Adverbial Clause <advcla>.

Special Rules: The prep tag (prepositional phrase) and todo tag (to-infinitive) do not nest within prep tags and todo tags but can nest other tags; wh tags (clause introducers) do not nest any other tags.

As a linguist expert, you are adept at analyzing sentences based on the aforementioned syntactic structure tagging system. Please refer to the provided examples and follow the established format to analyze the sentence I provide to you.

Sentence: Under these deals , the RTC sells just the deposits and the healthy assets .
Output: (sen (prep (w Under)(mod (w these))(w details)(w ,))(sbj (mod (w the))(w RTC))(prd (w sells))(obj (phrcoo (mod (w just))(mod (w the))(w deposits)(cc (w and))(mod (w the))(mod (w healthy))(w assets)(w .))))
Sentence: A Candian bank bought another thrift , in the first RTC transaction with a foreign bank .
Output: (sen (sbj (mod (w A))(mod (w Candian))(w bank))(prd (w bought))(obj (mod (w another))(w thrift)(w ,))(prep (w in)(mod (w the))(mod (w first))(mod (w RTC))(w transaction)(w with)(mod (w a))(mod (w foreign))(w bank)(w .)))
Sentence: Two of the four big thrifts were sold to NANB Crop. , Charlotte , N.C. , which has aggressively expanded its markets , particularly in Texas and Florida .
Output: (sen (sbj (w Two)(prep (w of)(mod (w the))(mod (w four))(mod (w big))(w thrifts)))(prd (w were)(w sold))(prep (w to)(w NCNB)(w Corp.)(w ,)(mod (w Charlotte)(w ,)(w N.C.)(w ,))(cla (wh (w which))(prd (w has)(mod (w aggressively))(w expanded))(obj (mod (w its))(w obj)(w ,))(mod (w particularly))(w in)(w Texas)(w and)(w Florida)(w .))))

Please analyze the given sentence: ...

Figure 3: Prompt in LLMs experiment.

# Cost-efficient Crowdsourcing for Span-based Sequence Labeling: Worker Selection and Data Augmentation

**Yujie Wang**[1,2*] **Chao Huang**[3*] **Liner Yang**[2,4†] **Zhixuan Fang**[5]
**Yaping Huang**[1] **Yang Liu**[2,4] **Jingsi Yu**[2,4] **Erhong Yang**[2,4]
[1]School of Computer and Information Technology, Beijing Jiaotong University, China
[2]National Language Resources Monitoring and Research Center for Print Media, Beijing Language and Culture University, China
[3]Department of Computer Science, The University of California, Davis, USA
[4]School of Information Science, Beijing Language and Culture University, China
[5]Institute for Interdisciplinary Information Sciences, Tsinghua University, China
lineryang@gmail.com

## Abstract

This paper introduces a novel crowdsourcing worker selection algorithm, enhancing annotation quality and reducing costs. Unlike previous studies targeting simpler tasks, this study contends with the complexities of label interdependencies in sequence labeling. The proposed algorithm utilizes a Combinatorial Multi-Armed Bandit (CMAB) approach for worker selection, and a cost-effective human feedback mechanism. The challenge of dealing with imbalanced and small-scale datasets, which hinders offline simulation of worker selection, is tackled using an innovative data augmentation method termed *shifting*, *expanding*, and *shrinking* (SES). Rigorous testing on CoNLL 2003 NER and Chinese OEI datasets showcased the algorithm's efficiency, with an increase in $F_1$ score up to 100.04% of the expert-only baseline, alongside cost savings up to 65.97%. The paper also encompasses a dataset-independent test emulating annotation evaluation through a Bernoulli distribution, which still led to an impressive 97.56% $F_1$ score of the expert baseline and 59.88% cost savings. Furthermore, our approach can be seamlessly integrated into Reinforcement Learning from Human Feedback (RLHF) systems, offering a cost-effective solution for obtaining human feedback. All resources, including source code and datasets, are available to the broader research community at https://github.com/blcuicall/nlp-crowdsourcing.

## 1 Introduction

Crowdsourcing, the practice of obtaining labeled data from a multitude of contributors (Howe, 2006), has emerged as a pivotal tool in data collection for deep learning models. It offers a cost-effective alternative to expert labeling, making it especially valuable in today's data-driven research landscape (Nowak and Rüger, 2010). While its application spans various domains, from image labeling to text classification (Venanzi et al., 2014), this paper narrows its focus on span-based sequence labeling tasks, which assign categorical labels to individual words within a sentence (Erdogan, 2010). Notable examples of such tasks include named entity recognition (NER) and opinion expression identification (OEI) (Collobert et al., 2011).

The inherent complexity of sequence labeling lies in the interdependencies of labels within a sequence. Unlike simpler tasks where labels are independent, sequence labeling requires contextual understanding, making it inherently more challenging (Rodrigues et al., 2014). Consequently, annotations from crowd workers, who might not possess the expertise of trained annotators, often exhibit reduced accuracy. This underscores the imperative to enhance annotation quality, a challenge that this study addresses.

[*]Equal contribution.
[†]Corresponding author: Liner Yang

©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License
This work was supported by the MOE (Ministry of Education in China) Project of Humanities and Social Sciences (No. 23YJCZH264) and the funds of Research Project of the National Language Commission (No. ZDA145-17).

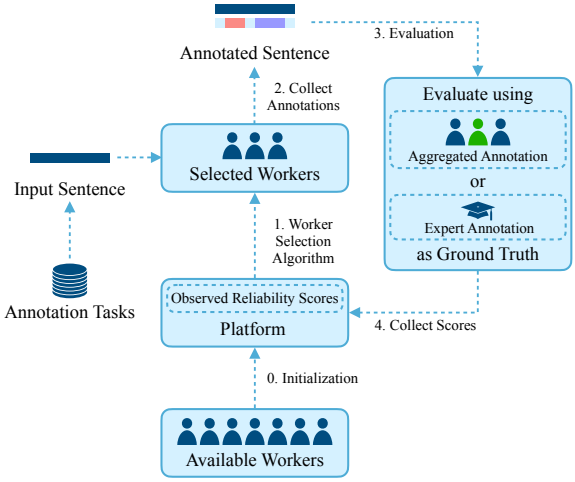



Figure 1: Our online worker selection framework for crowdsourcing.

A significant motivation driving this research is the potential application of a mixed feedback mechanism in Reinforcement Learning from Human Feedback (RLHF) systems. RLHF systems traditionally rely heavily on expert feedback, which, while accurate, is expensive and often not scalable (Casper et al., 2023). By integrating feedback from both experts and aggregated crowd workers, we can achieve a balance between accuracy and cost. This hybrid approach not only maintains the quality of feedback but also significantly reduces the financial burden, making RLHF systems more accessible and scalable.

Historically, research in this domain has concentrated on annotation aggregation (Rodrigues et al., 2014; Nguyen et al., 2017; Simpson and Gurevych, 2019), employing methods post data collection. However, given the varied skill levels among crowd workers, a proactive approach that identifies and leverages the most accurate workers during the data collection phase can significantly enhance data quality. Termed as online worker selection, this strategy involves iterative allocation of a set budget across a pool of workers to optimize annotation quality (Chen et al., 2013). This dynamic process grapples with the uncertainty of worker skill levels, necessitating a balance between exploring new workers and exploiting currently identified proficient ones.

In the context of sequence labeling, traditional bandit-based algorithms (Rangi and Franceschetti, 2018) fall short due to the intricacies introduced by label dependencies. These intricacies manifest challenges in both annotation evaluation and aggregation. To address the evaluation challenge, this study employs the span-level $F_1$ score (Derczynski, 2016), a widely recognized metric, as the feedback signal in the worker selection process. The core **challenge** here is the accurate computation of the $F_1$ score in the absence of expert annotations as a reference. The objective is to minimize reliance on costly expert annotations. For aggregation, while the majority voting method is employed for its simplicity and effectiveness, its reliability can be compromised when faced with divergent annotations from different annotators.

The overarching **goal** of this research is to maximize the quality of annotations while minimizing costs. This involves strategically replacing expert ground truth labels with aggregated crowd-sourced labels, ensuring that the overall $F_1$ score remains high. Such replacements are made only when there's a high level of agreement among crowd workers, indicating that expert evaluation might be redundant for that particular sequence. The proposed worker selection algorithm, as illustrated in Figure 1, adopts an iterative approach: tasks are assigned to a subset of workers, their annotations are evaluated, and the resulting scores inform worker selection in subsequent rounds.

However, real-world datasets present challenges due to their imbalanced nature and limited scale (Rodrigues et al., 2014; Zhang et al., 2022). Addressing these challenges, this paper introduces a data augmentation method tailored for span-based sequence labeling datasets. This method, designed to emulate potential human annotation errors, ensures that aggregated annotations remain meaningful. Three specific modifications, namely shifting, expanding, and shrinking, are applied to expert annotations,

generating a spectrum of potential human annotations. This augmentation addresses dataset limitations, facilitating the offline evaluation of worker selection algorithms.

In summary, this paper's contributions are manifold:

- It presents the exploration of worker selection for span-based sequence labeling tasks, recognizing the unique challenges they present.
- It employs the span-level $F_1$ score, evaluated by both experts and crowd workers, as a feedback mechanism, ensuring accurate worker selection.
- It introduces a data augmentation technique to counteract the limitations of real datasets, enabling effective offline simulations.
- Through rigorous experimentation, it demonstrates the efficacy of the proposed method, achieving impressive $F_1$ scores while significantly reducing expert annotation costs.

## 2 Related Work

Many studies (Rodrigues et al., 2014; Rodrigues and Pereira, 2018; Nangia et al., 2021) have used crowdsourcing for its efficiency and scalability. However, crowdsourcing suffers from the diversity of crowd workers' expertise and effort levels that are hardly measurable to task requesters. Different approaches to improving the quality of collected data have been proposed and studied. For span-based sequence labeling tasks, prior studies mainly focus on annotation aggregation. Rodrigues et al. (2014) proposed CRF-MA, a CRF-based model with an assumption that only one worker is correct for any label. HMM-crowd from Nguyen et al. (2017) outperforms CRF-MA, but the effect of sequential dependencies is not taken into account. Simpson and Gurevych (2019) uses a fully Bayesian approach BSC which is proved to be more effective in handling noise in crowdsourced data. Aggregation methods are used *after* the data collection process completes. But we aim to assure data quality and reduce cost *during* collecting. To this end, we focus on worker selection in our paper.

In online worker selection, we need to balance between exploring new workers and exploiting observed good workers. This exploration-exploitation tradeoff is extensively studied in the bandit literature (Lai and Robbins, 1985). In practice, we usually employ multiple crowd workers at the same time to finish the tasks more effectively. The combinatorial multi-armed bandit (CMAB) (Chen et al., 2013) models this circumstance. Biswas et al. (2015) and Rangi et al. (2018) reformulate the problem as a bounded knapsack problem (BKP) and address it with the B-KUBE (Tran-Thanh et al., 2014) algorithm. Song et al. (2021) introduce empirical entropy as the metric in CMAB and minimize the cumulative entropy with upper confidence bound (UCB) based algorithm. Li et al. (2022) consider the scalability of worker selection on large-scale crowdsourcing systems. These studies propose different methods under the CMAB settings, but on more complex span-based sequence labeling tasks there exists no discussion. We present the study of worker selection with CMAB on span-based sequence labeling tasks and show that our work performs well on the quality and efficiency of data collection.

## 3 Methodology

Consider an online crowdsourcing system that can reach out to a group of crowd workers $W = \{w_1, w_2, \ldots, w_N\}$. The workers are required to provide sequential annotations to a set of sentences $S = \{s_1, s_2, \ldots, s_M\}$. More specifically, a worker annotates a sentence by assigning a tag from a finite possible tag set $C$ (e.g., a set of BIO tags (Ramshaw and Marcus, 1995)) to each word. An annotation on sentence $s_i$ by worker $w_j$ is a tag sequence $\boldsymbol{a}_{ij} = a_1 a_2 \ldots a_k \ldots a_l$ where $a_k \in C$ and $l$ denotes the length of the sentence. We assume that every sentence is annotated by $K$ different workers independently. We define a task as the process of annotating one entire sentence, and hence there are in total $KM$ tasks. We seek to acquire an annotated dataset in which the average $F_1$ score of $\boldsymbol{a}_{ij}$ is maximized. If we know which workers give the best annotations a priori, we can simply ask these workers to finish all the tasks. However, such information is unavailable in practice, and we aim to design an algorithm that learns the best workers throughout the crowdsourcing process.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1239–1256, Taiyuan, China, July 25 – 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

In the beginning, we let each crowd worker annotate one sentence. We also ask the experts(e.g., well-trained linguists assumed to give the most precise annotations) to give one annotation for each of these sentences. Then we calculate the $F_1$ score of the annotation with the expert annotations as ground truth. We use these scores as the initial $F_1$ scores of workers. At each time step $t$ after initialization (as illustrated in Figure 1), we select a subset of workers $W_t \subset W$ to do annotation, based on criteria discussed in Section 3.2. The size of the subset $W_t$ should be neither too big nor too small (e.g., $0.3N$). We randomly choose a subset of sentences $S_t \subset S$, assign each $s_i \in S_t$ to $K$ different workers in $W_t$, and collect their annotations $\boldsymbol{A}_i = \{\boldsymbol{a}_{i1}, \boldsymbol{a}_{i2}, \ldots, \boldsymbol{a}_{iK}\}, \forall i \in \{1, 2, \ldots, |S_t|\}$. To evaluate workers' $F_1$ scores on $\boldsymbol{A}_i$, one can use the expert annotations as the ground truth, which, however, can be very expensive (İren and Bilgen, 2014). To cut down this cost, we reduce the usage of expert evaluations whenever crowd annotations are similar enough. We use the Fleiss' Kappa score $\kappa$ to measure this similarity. The $\kappa$ score ($\kappa \leq 1$) is a statistical measure of inter-annotator agreement. A larger value of $\kappa$ indicates stronger agreement between the workers. $\kappa$ score exceeding an empirical threshold indicates that the crowd workers reach a consensus on $s_i$. In that case, we aggregate $\boldsymbol{A}_i$ with MV and use the aggregated annotation as the ground truth of sentence $s_i$. If the workers do not reach a consensus, we resort to expert annotations as ground truth. Next, we can calculate the $F_1$ scores of each $\boldsymbol{a}_{ij} \in \boldsymbol{A}_i$ and update the $F_1$ scores of the selected workers.

### 3.1 Problem Formulation

At time $t$, we obtain $K$ crowd annotations $\boldsymbol{A}_i$ on each sentence $s_i \in S_t$. We denote all annotations collected on $S_t$ by $\boldsymbol{\mathcal{A}}_t = \{\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots, \boldsymbol{A}_{|S_t|}\}$. To simplify our expression, we use $\mathrm{F}_1^{\mathrm{Exp}}(\boldsymbol{a}_{ij})$ to represent the $F_1$ score of $\boldsymbol{a}_{ij}$ using expert annotation as ground truth, and $\mathrm{F}_1^{\mathrm{MV}}(\boldsymbol{a}_{ij})$ to represent the $F_1$ score of $\boldsymbol{a}_{ij}$ using the MV aggregation of $\boldsymbol{A}_i \in \boldsymbol{\mathcal{A}}_t$ as ground truth. On collected annotation sets, $\mathrm{F}_1^{\mathrm{Exp}}(\boldsymbol{A}_i)$ denotes the average $F_1$ score of all $\boldsymbol{a}_{ij} \in \boldsymbol{A}_i$. Similarly, $\mathrm{F}_1^{\mathrm{Exp}}(\boldsymbol{\mathcal{A}}_t)$ denotes the average $F_1$ score of all $\boldsymbol{A}_i \in \boldsymbol{\mathcal{A}}_t$. As $\mathrm{F}_1^{\mathrm{Exp}}(\boldsymbol{\mathcal{A}}_t)$ reflects the true accuracy of crowd annotations, our objective is to maximize the average expectation, or equivalently the cumulative expectation of $\mathrm{F}_1^{\mathrm{Exp}}(\boldsymbol{\mathcal{A}}_t)$ over time $T$. We formulate this problem as a CMAB problem below:

$$\max \quad \sum_{t=1}^{T} \mathbb{E}[\mathrm{F}_1^{\mathrm{Exp}}(\boldsymbol{\mathcal{A}}_t)] \tag{1}$$

$$\text{s.t.} \quad W_t \subset W, t \in \{1, 2, \ldots, T\} \tag{2}$$

Since we have no information about workers' average $F_1$ scores, we need to balance exploring potentially better workers and exploiting the current best workers during worker selection. This tradeoff is extensively discussed in bandit literature where arms with unknown distributions form super-arms. The arms are associated with a set of random variables $X_{j,t}$ with bounded support on [0, 1]. Variable $X_{j,t}$ indicates the random outcome of arm $j$ in time step $t$. The set of random variables $\{X_{j,t} | t \geq 1\}$ associated with arm $j$ are independent and identically distributed according to certain unknown distribution $D_j$ with unknown expectation $\bar{\mu}_j$. The platform plays a super-arm at each time step, and the reward of arms in it is revealed. These rewards are used as a metric for selecting the super-arm in future time steps. After enough time steps, the platform will be able to identify the best super-arm and keep playing it to maximize the overall reward. Similar to bandit terminologies, we call each worker $w_j \in W$ an arm and the worker subset $W_t \subset W$ a super-arm selected at $t$.

### 3.2 Worker Selection Algorithm

Specifically, there are three methods to calculate the reward of worker $w_j$ at time step $t$ as follows.

**Expert Only** This is a benchmark approach where the $F_1$ score is calculated using only expert annotations as ground truth. This method provides intuitively the most accurate $F_1$ scores. The reward of worker $w_j$ is defined as:

$$\mu_j^{\mathrm{Exp}}(t) = \mathrm{F}_1^{\mathrm{Exp}}(\boldsymbol{a}_{ij}(t)) \tag{3}$$

CCL 2024

The expert-only method requires an expert annotation on every sentence, which is costly and usually not practical.

**Majority Voting (MV)** To reduce expert annotations, we aggregate $\boldsymbol{A}_i$ for each sentence $s_i$, and use the aggregated annotation via MV as ground truth, i.e.,

$$\mu_j^{\text{MV}}(t) = \text{F}_1^{\text{MV}}(\boldsymbol{a}_{ij}(t)) \tag{4}$$

**Expert+MV** When workers give very different annotations on the same sentence (usually when the task is difficult), one can be uncertain about the voted (and possibly noisy) ground truth. In this case, we want to resort to both crowd workers and experts. The choice is based on the well-known Fleiss' Kappa score $\kappa$ that can quantitatively evaluate the agreement of crowd workers. For each sentence $s_i$, if $\kappa(\boldsymbol{A}_i)$ is greater than a preset empirical threshold value $\tau$, the reward of annotating workers is $\text{F}_1^{\text{MV}}(\boldsymbol{a}_{ij}(t))$. Otherwise, the reward is $\text{F}_1^{\text{Exp}}(\boldsymbol{a}_{ij}(t))$. In this way, MV is only used when the crowd workers can reach an agreement. Thus the reward is always calculated based on reliable ground truth. We summarize the reward of worker $w_j$ as:

$$\mu_j^{\text{Exp+MV}}(t) = \begin{cases} \text{F}_1^{\text{MV}}(\boldsymbol{a}_{ij}(t)), & \kappa(\boldsymbol{A}_i) > \tau \\ \text{F}_1^{\text{Exp}}(\boldsymbol{a}_{ij}(t)), & \kappa(\boldsymbol{A}_i) \leq \tau \end{cases} \tag{5}$$

The $\epsilon$-Greedy, Thompson Sampling, and Combinatorial Upper Confidence Bound (CUCB) are three effective algorithms to solve the CMAB problem. For each worker $w_j \in W$, both algorithms maintain a variable $\bar{\mu}_j(t)$ as the average reward (i.e., the average $\text{F}_1$ score) of worker $w_j$ at time step $t$. CUCB additionally maintains a variable $T_j(t)$ as the total number of sentences worker $w_j$ has annotated till time step $t$. Details of the worker selection algorithm with our **Exp.+MV** metric are shown in Algorithm 1. As for the selection criterion mentioned in the algorithm, $\epsilon$-Greedy utilize a hyper-parameter $\epsilon$ which refers to the probability of exploring random workers. Thus $1-\epsilon$ refers to the probability of exploiting the best workers till the current time step. Formally, $W_t$ is selected with a random variable $p \in [0, 1]$ as below:

$$W_t = \begin{cases} \text{random } W_t \subset W, & p < \epsilon \\ \underset{W_t \subset W}{\text{argmax}} \sum_{w_j \in W_t} \bar{\mu}_j, & p \geq \epsilon \end{cases} \tag{6}$$

Thompson Sampling samples from gaussian distributions of workers' rewards at each time step $t$, and select workers which could maximize the total reward. CUCB handles the tradeoff by adding an item considering $T_j$ and $t$ to $\bar{\mu}_j$ like:

$$W_t = \underset{W_t \subset W}{\text{argmax}} \sum_{w_j \in W_t} \left( \bar{\mu}_j + \sqrt{\frac{3 \ln t}{2 T_j}} \right) \tag{7}$$

This makes workers with less annotations more likely to be selected as the algorithm proceeds. We provide a brief analysis in Appendix B. We explain on the application of our worker selection algorithms when building new datasets in Appendix E

### 3.3 Data Augmentation Method

We propose the data augmentation method to facilitate the offline simulation of the crowdsourcing process, thus evaluating the worker selection algorithms. During offline simulation, when the worker selection strategy selects a certain worker to annotate a certain sentence, we can use the annotation in the original dataset if it exists. But if the selected worker did not annotate the sentence in the original dataset, we need to generate an annotation for the sentence. And the generated annotation should be in the same quality (depicted in F-score) as the real annotations by the worker. The generated annotation will be then used with the other annotations on the same sentence for majority voting.

Generating the missing annotations for each worker $w_j$ is a great challenge when we expect the generated annotations to reflect the factual reliability of $w_j$. In other words, we expect the average $\text{F}_1$ score

CCL 2024

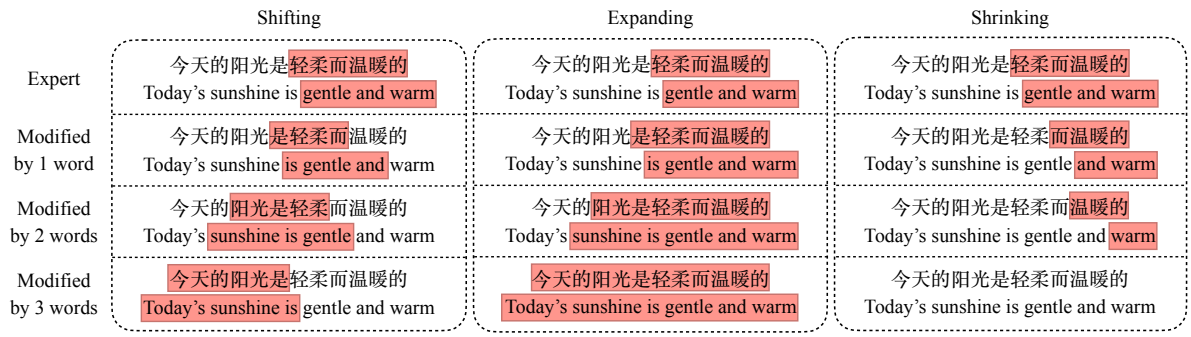



Figure 2: An example of the three methods to generate annotations. Chinese characters and corresponding English words with red backgrounds indicate annotation spans.

of each $w_j \in W$ to remain constant before and after augmenting the dataset with generated annotations. This is critical and difficult since real datasets are imbalanced and of small scale that cannot well support worker selection algorithms.

As there lack previous work on generating missing crowd annotations for span-based sequence labeling, we start with several naive algorithms such as randomly generating label sequences as annotations, and mixing expert annotations with completely incorrect (e.g., empty) annotations. But these algorithms either cannot produce annotations with expected $F_1$ scores, or generate confusing annotations which make later aggregation meaningless. This motivates us to design a data augmentation method specialized for span-based sequence labeling datasets.

Through our statistical analysis and observation on the real datasets, we characterized the 3 most common annotation error patterns. Due to space limitation, we defer the detailed analysis to Appendix C. Based on these analysis results, we propose a data augmentation method as follows: For each sentence $s_i \in S$, we modify the annotation span based on the expert annotation. We use three types of modifications to generate new annotation spans with different $F_1$ scores as illustrated in Figure 2. The goal of these modifications is to simulate varying annotation errors made by human annotators.

**Shifting** We move both the left and the right border of the annotation span simultaneously in the same direction by one word per step.

**Expanding** We set one of the span borders fixed, and move the other border by one word per step to *increase* the length of the annotation span.

**Shrinking** We set one of the span borders fixed, and move the other border by one word per step to *decrease* the length of the annotation span.

We perform these modifications on a span multiple times, generating new annotation spans, until (1)the modified span does not overlap with the original one, (2)one of the span borders reaches an end of sentence or another span in the same sentence, or (3) the span length becomes 0.

For each sentence $s_i \in S$, $s_i$ may contain multiple annotation spans. We perform modifications on each span in $s_i$, and find all combinations of spans to form possible sentence annotations. With these methods, we can imitate crowd annotations with different kinds of errors in practice. Next, for each worker $w_j \in W_{ti}$, if $w_j$ has no annotation on $s_i$ in the original dataset, we select one from all the expert and generated annotations on $s_i$.

We first calculate $\bar{\varphi}_j$ as the average $F_1$ score of all annotations by $w_j$ on the original dataset, and then follow the detailed steps described in Algorithm 2 to do the selection. We aim to keep the overall $F_1$ score of $w_j$ unchanged.

To better illustrate the procedure of the augmentation, we provide a running example in Appendix D.

CCL 2024

| Measure | Chinese OEI | CoNLL 2003 |
|---|---|---|
| # of Sentence | 8047 | 4580 |
| # of Worker | 70 | 47 |
| Span Length | 5.05 | 1.51 |
| Max | 658 | 1626 |
| Min | 153 | 48 |
| Range | 505 | 1578 |
| Mean | 368 | 350 |
| Median | 332.5 | 230 |
| SD | 135.23 | 328.01 |
| Variance | 18286.52 | 107589.34 |
| CV | 36.71% | 93.57% |

Table 1: Statistics of the original datasets. Span lengths are averages. The terms SD and CV represent Standard Deviation and Coefficient of Variation respectively. The metrics Max, Min, Range, Mean, Median, SD, Variance, and CV pertain to the number of sentences annotated by each worker, indicating dataset imbalances.

| Worker ID | Rnd. Gen. $\lvert\Delta F_1\rvert$ | SES Only $\lvert\Delta F_1\rvert$ | SES +Alg.2 $\lvert\Delta F_1\rvert$ |
|---|---|---|---|
| 25 | 2.83 | 6.69 | 0.01 |
| 52 | 8.15 | 10.83 | 0.00 |
| 46 | 3.83 | 13.48 | 0.00 |
| 43 | 10.02 | 11.21 | 0.00 |
| 18 | 9.87 | 12.84 | 0.00 |
| 50 | 16.69 | 10.71 | 0.00 |
| 12 | 47.18 | 10.52 | 0.00 |
| *Avg.* | 14.08 | 10.90 | 0.0014 |

Table 2: Comparisons between different data augmentation methods on the error of span-level exact $F_1$ score of every crowd worker. The error $|\Delta F_1|$ is calculated as the absolute difference between each worker's $F_1$ score after augmentation and his real $F_1$ score. The methods **Rnd. Gen.**, **SES Only** and **SES + Alg.2** are introduced in Section 3.3.

## 4 Experiments

### 4.1 Original Datasets

We compare our CMAB-based algorithms to several widely adopted baselines on two span-based sequence labeling datasets.

**CoNLL 2003** The CoNLL 2003 English named-entity recognition dataset (Tjong Kim Sang and De Meulder, 2003) is a collection of news article from Reuters Corpus (Lewis et al., 2004). The dataset contains only expert annotations for four named entity categories (PER, LOC, ORG, MISC). Rodrigues et al. (2014) collected crowd annotations on 400 articles from the original dataset.

**Chinese OEI** The Chinese OEI dataset (Zhang et al., 2022) consists of sentences on the topic of COVID-19 collected from Sina Weibo[1], in which the task is to mark the spans of opinion expressions. The Chinese OEI dataset contains expert and crowd labels for two opinion expression categories (POS, NEG). Detailed statistics are shown in Table 1.

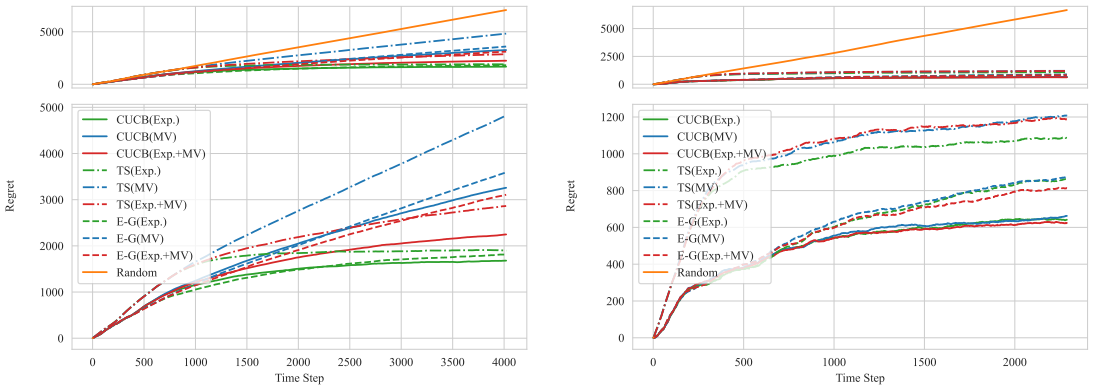
### 4.2 Data Augmentation

We augment both datasets with the method proposed in Section 3.3. According to Table 1, the most hard-working annotator in the OEI dataset provided annotations on 658 sentences, while the least one annotated only 153 sentences. On average, each crowd worker annotated 368 out of 8047 sentences in the Chinese OEI dataset. For the offline simulation of the worker selection process, we want every worker to annotate all 8047 sentences. Therefore we need to generate the missing 8047 - 368 = 7679 annotations for every worker, on average. This also applies similarly to the CoNLL 2003 dataset.

Through our method, the average $F_1$ score of each $w \in W$ remains nearly unchanged before and after augmenting the original dataset with generated annotations[2]. Due to space limitation, we present comparisons of different augmentation algorithms with 10 sampled workers in Table 2. The complete results are deferred to Table 7 in the appendix. These results show that our **SES + Alg.2** method clearly

[1] https://english.sina.com/weibo/

[2] The augmentation procedure takes about 2 hours on a computer with a 2.9 GHz Quad-Core Intel Core i7 CPU.




(a) Chinese OEI (b) CoNLL 2003

Figure 3: Cumulative regrets w.r.t time steps of all different worker selection methods.

| Method | Token-level | | | Span-level Exact | | | Span-level Prop. | | |
|---|---|---|---|---|---|---|---|---|---|
| | P | R | $F_1$ | P | R | $F_1$ | P | R | $F_1$ |
| Oracle | 65.69 | 83.99 | 70.00 | 78.15 | 72.23 | 74.96 | 87.97 | 80.03 | 83.82 |
| Random | 55.95 | 66.42 | 57.50 | 64.42 | 55.64 | 59.40 | 75.70 | 62.61 | 68.54 |
| $\epsilon$-G (Exp.) | 64.94 | 80.48 | **68.56** | 75.24 | 68.16 | **71.34** | 85.85 | 76.79 | **81.06** |
| $\epsilon$-G (MV) | 64.44 | 80.22 | 67.98 | 74.69 | 67.59 | 70.77 | 85.67 | 76.09 | 80.59 |
| $\epsilon$-G (Exp.+MV) | 64.68 | 80.94 | **68.41** | 75.08 | 68.37 | **71.40** | 85.93 | 76.62 | **81.01** |
| TS (Exp.) | 64.94 | 79.88 | **68.51** | 75.64 | 68.31 | **71.57** | 85.02 | 75.71 | **80.09** |
| TS (MV) | 64.47 | 79.19 | 67.91 | 74.97 | 67.54 | 70.80 | 84.14 | 74.21 | 78.86 |
| TS (Exp.+MV) | 64.20 | 79.09 | **67.62** | 75.27 | 67.83 | **71.12** | 84.77 | 75.39 | **79.81** |
| CUCB (Exp.) | 65.65 | 80.34 | **69.24** | 75.94 | 69.12 | **72.20** | 86.17 | 77.22 | **81.45** |
| CUCB (MV) | 65.39 | 80.00 | 68.91 | 75.95 | 68.90 | 72.08 | 86.13 | 76.67 | 81.12 |
| CUCB (Exp.+MV) | 65.33 | 81.12 | **69.11** | 75.70 | 69.30 | **72.21** | 86.17 | 77.28 | **81.48** |

Table 3: Detailed P, R, and $F_1$ scores of all methods on the CoNLL 2003 dataset. All our algorithms perform significantly better than the Random (i.e., naive crowdsourcing) baseline.

outperforms the other baselines, producing almost the same $F_1$ scores for each worker as their original ones.

### 4.3 Worker Selection

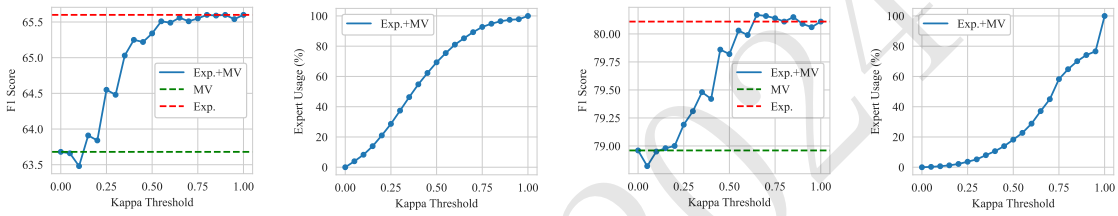
**Baselines** We test the **Exp.+MV** method with 4 baselines: **Oracle**, **Random**, **Exp.**, and **MV**. **Oracle** always selects the empirical best super-arm $W^{opt}$ at every time step $t$. **Random** selects a different set of workers randomly at every time step $t$, which is equivalent to usual crowdsourcing procedure without worker selection. **Exp.**, **MV**, and **Exp.+MV** are CMAB-based algorithms introduced in Section 3.2. The CMAB-based algorithms are tested with CUCB, Thompson Sampling and $\epsilon$-Greedy as the worker selection criterion respectively.

**Regret as a Metric** We evaluate our worker selection algorithms using cumulative regret, a metric indicating the performance deviation from the oracle's selection defined as:

$$R(T) = \sum_{t=1}^{T} \left( \sum_{w_j \in W^{opt}} \bar{\mu}_j - \sum_{w_k \in W_t} \mu_k(t) \right) \tag{8}$$

| Method | Token-level | | | Span-level Exact | | | Span-level Prop. | | |
|---|---|---|---|---|---|---|---|---|---|
| | P | R | $F_1$ | P | R | $F_1$ | P | R | $F_1$ |
| Oracle | 62.88 | 68.62 | 64.80 | 54.48 | 51.97 | 53.07 | 72.79 | 64.07 | 68.15 |
| Random | 58.49 | 57.30 | 57.42 | 43.99 | 35.50 | 39.18 | 69.01 | 52.36 | 59.55 |
| $\epsilon$-G (Exp.) | 61.91 | 64.58 | **62.61** | 51.72 | 46.37 | **48.76** | 72.28 | 60.25 | **65.72** |
| $\epsilon$-G (MV) | 60.87 | 63.52 | 61.55 | 48.72 | 44.66 | 46.37 | 70.15 | 58.94 | 64.05 |
| $\epsilon$-G (Exp.+MV) | 61.76 | 64.46 | **62.47** | 49.14 | 45.35 | **46.96** | 71.21 | 59.92 | **65.08** |
| TS (Exp.) | 62.66 | 64.91 | **63.20** | 49.76 | 42.34 | **45.69** | 72.15 | 60.20 | **65.63** |
| TS (MV) | 59.82 | 61.90 | 60.25 | 44.81 | 40.71 | 42.36 | 67.72 | 56.05 | 61.34 |
| TS (Exp.+MV) | 61.66 | 64.03 | **62.23** | 47.20 | 42.36 | **44.49** | 70.66 | 59.07 | **64.35** |
| CUCB (Exp.) | 63.02 | 63.75 | **62.93** | 52.24 | 45.51 | **48.56** | 73.05 | 59.53 | **65.60** |
| CUCB (MV) | 61.94 | 62.09 | 61.55 | 49.57 | 44.39 | 46.66 | 71.22 | 57.59 | 63.68 |
| CUCB (Exp.+MV) | 62.83 | 63.62 | **62.75** | 51.31 | 45.60 | **48.16** | 72.48 | 59.33 | **65.25** |

Table 4: Detailed P, R, and $F_1$ scores of all methods on the Chinese OEI dataset. All our algorithms perform significantly better than the Random (i.e., naive crowdsourcing) baseline.




(a) $F_1$ score w.r.t $\tau$ on the Chinese OEI dataset. (b) Expert usage w.r.t $\tau$ on the Chinese OEI dataset. (c) $F_1$ score w.r.t $\tau$ on the CoNLL 2003 dataset. (d) Expert usage w.r.t $\tau$ on the CoNLL 2003 dataset.

Figure 4: $F_1$ scores of the produced annotations and usage of expert for annotation evaluations w.r.t the kappa threshold $\tau$ of the **Exp.+MV** method on Chinese OEI and CoNLL 2003 datasets.

In our experiments, we request 10 annotations per sentence, allowing CMAB-based algorithms to converge, and select 20 workers at each time step $t$. On the Chinese OEI dataset, setting the kappa threshold $\tau$ to 0.4 in **Exp.+MV** results in a 57.02% reduction in expert annotation cost, while a 0.65 threshold on the CoNLL 2003 dataset leads to a 43.83% cost reduction.

Results show **Random** consistently underperforms across datasets. On the Chinese OEI dataset, **Exp.+MV** surpasses **MV**, albeit with higher regret than **Exp.**, justified by the substantial cost savings. On the CoNLL 2003 dataset, **Exp.+MV** even outperforms **Exp.**, suggesting crowd workers can provide valuable input for simpler tasks like NER. Overall, algorithms employing the CUCB criterion demonstrate superior performance, with **CUCB (Exp.+MV)** excelling in balancing cumulative regret and expert cost.

**Effect of $\tau$ on $F_1$ and cost** Next, we discuss how different kappa threshold values $\tau$ affect the average $F_1$ score of the produced annotation dataset. We test $\tau \in [0, 1]$ with a step of 0.05. In real datasets like CoNLL 2003 and Chinese OEI, the number of annotations per sentence is often quite small. To better fit the practical situations, we ask for 4 annotations on each sentence in the following experiments. Other settings remain unchanged. Since CUCB performs better than Thompson Sampling and $\epsilon$-Greedy on both datasets, we display only the results from CUCB in later experiments.

On the Chinese OEI dataset, as illustrated in Figure 4(a) and 4(b), $F_1$ increases sharply with $\tau \in [0, 0.4]$. When $\tau = 0.4$, **Exp.+MV** achieves 99.47% $F_1$ score of **Exp.**, and saves 47.19% of the expert cost. The $F_1$ score goes up slowly until $\tau$ reaches 0.8. When $\tau = 0.8$, the $F_1$ score of **Exp.+MV** becomes

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1239–1256, Taiyuan, China, July 25 – 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

exactly the same as the one of **Exp.**, and **Exp.+MV** still saves 6.6% of the expert cost.

The results on the CoNLL 2003 dataset are shown in Figure 4(c) and 4(d). Similarly, the $F_1$ score of the produced annotation dataset grows fast as $\tau \in [0, 0.45]$. When $\tau = 0.45$, the **Exp.+MV** method already produce an annotation dataset with its $F_1$ reaching 99.86% of **Exp.**. At this point, **Exp.+MV** saves 88.57% of the expert cost. When $\tau = 0.65$, **Exp.+MV** outperforms **Exp.** with a 100.04% $F_1$ score and a 65.97% reduction in expert usage.

Our **CUCB (Exp.+MV)** worker selection algorithm eliminates the need for expert evaluation on every sentence. Instead, we harness crowd intelligence via our kappa-thresholded MV, producing datasets of comparable or even superior quality to those relying solely on expert evaluations.

**Extended $F_1$ Metrics** All of the $F_1$ scores in the previous experiments are span-level proportional scores calculated by the proportion of the overlap referring to the expert annotation (Zhang et al., 2022). To provide additional comparisons between different methods, we also invoke token-level and span-level exact P, R, $F_1$ scores as supporting metrics. We run the whole process from data augmentation to worker selection with all 3 metrics separately. The kappa threshold $\tau$ in **Exp.+MV** is set to 0.4 on the Chinese OEI dataset and 0.65 on the CoNLL 2003 dataset. Detailed scores are listed in Table 3 and 4. The results show that **Exp.+MV** achieves scores as good as **Exp.** and much better than **MV**, which validates previous experiments and shows our worker selection methods are robust to different metrics.

**Feedback Simulator** We also test our worker selection methods with a feedback simulator. The simulator generates numerical feedback from $Bernoulli$ distribution in annotation evaluations. This is to eliminate the varying level of difficulty in different tasks and evaluate our worker selection algorithms under more stable settings. Our algorithm achieves good results on the simulator. We put the definitions and results in Appendix A.

**Effect on ML Models** To further show the effect of our worker selection algorithm on the performance of machine learning models, we have run experiments with several widely-accepted models and provide the results in Table 6. We observe a consistent increment of F1 score on the ML models, with our bandit-based worker selection algorithm. This validates that our worker selection algorithm may help improve the performance of ML models while saving budget on data crowdsourcing.

## 5 Conclusion

In this study, we introduced a CMAB-based worker selection strategy tailored for span-based sequence labeling tasks, leveraging the span-level $F_1$ with **Exp.+MV** as a feedback mechanism. To address the challenges posed by unbalanced and limited real datasets, we innovated a data augmentation method. This technique not only facilitates offline simulation but also mirrors the genuine annotation behaviors of workers closely.

Our empirical evaluations underscore the efficacy of the proposed method. On the Chinese OEI dataset, our approach achieved an impressive 99.47% $F_1$ score, translating to a substantial 47.19% reduction in expert costs. Similarly, on the CoNLL 2003 dataset, we observed a remarkable 100.04% $F_1$ score, with savings of up to 65.97% in expert costs, both benchmarks set against expert-evaluation-only baselines. Furthermore, our method demonstrated its robustness with a 94.86% $F_1$ score and a 65.97% reduction in expert costs on a data-free simulator. Our approach also boosts ML model performance, optimizing both accuracy and cost.

## References

Arpita Biswas, Shweta Jain, Debmalya Mandal, and Y. Narahari. 2015. A truthful budget feasible multi-armed bandit mechanism for crowdsourcing time critical tasks. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, pages 1101–1109.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. 2023. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*.

Wei Chen, Yajun Wang, and Yang Yuan. 2013. Combinatorial multi-armed bandit: General framework and applications. In *Proceedings of the 30th International Conference on Machine Learning*, pages 151–159.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(76):2493–2537.

Leon Derczynski. 2016. Complementarity, F-score, and NLP evaluation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 261–266.

Hakan Erdogan. 2010. Sequence labeling: Generative and discriminative approaches. In *Proceedings of the 9th International Conference on Machine Learning and Applications*, pages 1–132.

Evrard Garcelon, Vashist Avadhanula, Alessandro Lazaric, and Matteo Pirotta. 2022. Top k ranking for multi-armed bandit with noisy evaluations. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pages 6242–6269.

Jeff Howe. 2006. The rise of crowdsourcing. *Wired magazine*, 14(6):1–4.

Tze Leung Lai and Herbert Robbins. 1985. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathmatics*, 6(1):4–22.

David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. 2004. RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397.

Feng Li, Jichao Zhao, Dongxiao Yu, Xiuzhen Cheng, and Weifeng Lv. 2022. Harnessing context for budget-limited crowdsensing with massive uncertain workers. *IEEE/ACM Transactions on Networking*, 30(5):2231–2245.

Rajeev Motwani and Prabhakar Raghavan. 1995. *Randomized Algorithms*.

Nikita Nangia, Saku Sugawara, Harsh Trivedi, Alex Warstadt, Clara Vania, and Samuel R. Bowman. 2021. What ingredients make for an effective crowdsourcing protocol for difficult NLU data collection tasks? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 1221–1235.

An Thanh Nguyen, Byron Wallace, Junyi Jessy Li, Ani Nenkova, and Matthew Lease. 2017. Aggregating and predicting sequence labels from crowd annotations. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 299–309.

Stefanie Nowak and Stefan Rüger. 2010. How reliable are annotations via crowdsourcing: A study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the International Conference on Multimedia Information Retrieval*, page 557–566.

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

Anshuka Rangi and Massimo Franceschetti. 2018. Multi-armed bandit algorithms for crowdsourcing systems with online estimation of workers' ability. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 1345–1352.

Filipe Rodrigues and Francisco C. Pereira. 2018. Deep learning from crowds. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, pages 1611–1618.

Filipe Rodrigues, Francisco Pereira, and Bernardete Ribeiro. 2014. Sequence labeling with multiple annotators. *Machine learning*, 95(2):165–181.

Edwin Simpson and Iryna Gurevych. 2019. A Bayesian approach for sequence tagging with crowds. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 1093–1104.

Yiwen Song and Haiming Jin. 2021. Minimizing entropy for crowdsourcing with combinatorial multi-armed bandit. In *IEEE Conference on Computer Communications*, pages 1–10.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning*, pages 142–147.

CCL 2024

| **Method** | $F_1$ |
|---|---|
| Oracle | 74.12 |
| Random | 65.12 |
| Exp. | **69.78** |
| MV | 66.80 |
| Exp.+MV | **68.29** |

Table 5: The overall span-level proportional $F_1$ scores of all methods with the feedback simulator.

| **Method** | Original | w/ Our Alg. |
|---|---|---|
| LSTM-Crowd-cat | 52.66 | 54.27 |
| Bert-BiLSTM-CRF | 52.14 | 54.51 |
| Annotator-Adaptor | 53.86 | 56.16 |

Table 6: Span-level exact $F_1$ scores of widely-accepted deep learning models on the Chinese OEI dataset. LSTM-Crowd-cat is from Nguyen et al. (2017). Bert-BiLSTM-CRF and Annotator-Adaptor are from Zhang et al. (2022). We provide results with and without our worker selection algorithm.

Long Tran-Thanh, Sebastian Stein, Alex Rogers, and Nicholas R. Jennings. 2014. Efficient crowdsourcing of unknown experts using bounded multi-armed bandits. *Artificial Intelligence*, 214:89–111.

Matteo Venanzi, John Guiver, Gabriella Kazai, Pushmeet Kohli, and Milad Shokouhi. 2014. Community-based bayesian aggregation models for crowdsourcing. In *Proceedings of the 23rd International Conference on World Wide Web*, page 155–164.

Xin Zhang, Guangwei Xu, Yueheng Sun, Meishan Zhang, Xiaobin Wang, and Min Zhang. 2022. Identifying Chinese opinion expressions with extremely-noisy crowdsourcing annotations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 2801–2813.

Deniz İren and Semih Bilgen. 2014. Cost of quality in crowdsourcing. *Human Computation*, 1, 12.

## A Feedback Simulator

The performance of crowd workers can vary across different types of annotation tasks. To evaluate the **Exp.+MV** worker selection method in more stable conditions without task-specific influence, we do not actually annotate the sentences, but directly use a worker's average $F_1$ score to simulate his score on each sentence he annotates. The simulated scores are used as the numerical feedback for worker selection. Specifically, for each worker $w$, we calculate in advance two average $F_1$ scores for all of their annotations on the original dataset. The two $F_1$ scores for each worker are calculated using expert and majority vote (MV) evaluation respectively, denoted as $\bar{\varphi}_w^{Exp.}$ and $\bar{\varphi}_w^{MV}$. At each time step $t$, for every sentence $s_i$ in the sentence set to be annotated $S_t$, we ask $K$ different workers from the current selected workers $W_t$ to annotate it. Then, we use a random value between 0 and 1 as the agreement level $\kappa$. If $\kappa$ exceeds the threshold value $\tau$ (set to 0.4 in **Exp.+MV**), we independently generate feedback for the $K$ workers from a Bernoulli distribution with a probability parameter set to $\bar{\varphi}_w^{MV}$. If not, the feedback is generated from a Bernoulli distribution with a probability parameter set to $\bar{\varphi}_w^{Exp.}$. The span-level average $F_1$ scores of the annotated dataset using different worker selection algorithm are shown in Table 5. Our feedback mechanism **Exp.+MV** for worker selection achieved comparable performance to the expert-only mechanism **Exp.** (68.29 versus 69.78), while in the same time reduced expert involvement in evaluation by 59.88% under the dataset-independent conditions.

## B Regret Analysis

We provide a brief regret analysis of the worker selection framework assuming that we use the $\epsilon$-greedy algorithm and that each worker's reward follows a Bernoulli distribution.

**Algorithm 1** The worker selection algorithm with the Expert+MV metric.

Let each worker $w_j \in W$ annotate a random sentence and initialize variable $\bar{\mu}_j$ with $\mathrm{F}_1$ by expert evaluation
For each worker $w_j \in W$, initialize $T_j \leftarrow 1$
$t \leftarrow |W|$
**while** unannotated sentences exist **do**
  $t \leftarrow t + 1$
  Select $W_t \subset W$ based on certain criterion (e.g., (6), (7))
  Split $W_t$ into several disjoint subsets $\{W_{t1}, \ldots, W_{ti}, \ldots, W_{tn}\}$, each containing $K$ workers
  **for all** $W_{ti}$ **do**
    Let each $w_j \in W_{ti}$ annotate an sentence $s_i$ and collect the annotations $\boldsymbol{A}_i$
    **if** $\kappa(\boldsymbol{A}_i) > \tau$ **then**
      Update $T_j$ and $\bar{\mu}_j$ with $\mathrm{F}_1^{\mathrm{MV}}(\boldsymbol{a}_{ij}(t))$
    **else**
      Update $T_i$ and $\bar{\mu}_j$ with $\mathrm{F}_1^{\mathrm{Exp}}(\boldsymbol{a}_{ij}(t))$
    **end if**
  **end for**
**end while**

The main proof follows the proof of Theorem 1 in (Garcelon et al., 2022). The key contribution here is that we need to specify that the evaluation signal (generated by majority voting) is a generalized linear model of workers' true reward signal (generated by expert/oracle). To this end, we utilize the following form of the Chernoff bound which applies for any random variables with bounded support.

**Lemma 1** *(Chernoff Bound (Motwani and Raghavan, 1995)) Let $X_1, X_2, \cdots, X_N$ be independent random variables such that $x_l \leq X_i \leq x_h$ for all $i \in \{1, 2, \cdots, N\}$. Let $X = \sum_{i=1}^{N} X_i$ and $\mu = \mathbb{E}(X)$. Given any $\delta > 0$, we have the following result:*

$$P(X \leq (1-\delta)\mu) \leq e^{-\frac{\delta^2 \mu^2}{N(x_h - x_l)^2}}. \tag{9}$$

For the purpose of our discussion, let $X_i \in \{0, 1\}$ be a binary random variable, where $X_i = 0$ denotes that worker $i$ provides an incorrect solution, and $X_i = 1$ denotes that worker $i$ generates a correct solution. Define $X = \sum_{i \in \mathcal{N}} X_i$.

We aim to approximate $P_{\mathrm{MV}}$, which is the probability that the majority of the $N$ workers provide the correct estimate. We apply the Chernoff Bound in Lemma 1 to $P_{\mathrm{MV}}$. We can compute

$$\mathbb{E}(X) = \bar{p} = \frac{\sum_{i=1}^{N} p_i}{N}. \tag{10}$$

Based on (9), we let $\mu = \mathbb{E}(X),\ \delta = \frac{N(\bar{p}-\frac{1}{2})}{\frac{N}{2}+N(\bar{p}-\frac{1}{2})}$, $x_l = 0$, $x_h = 1$, and get the following result:

$$\begin{aligned}
P_{\mathrm{MV}} &= P\left(X \geq \frac{N}{2}\right) = 1 - P\left(X \leq \frac{N}{2}\right) \\
&\geq 1 - e^{-\frac{\delta^2 \mu^2}{N}} && (11) \\
&= 1 - e^{-\frac{\frac{N^2(\bar{p}-\frac{1}{2})^2}{[\frac{N}{2}+N(\bar{p}-\frac{1}{2})]^2}[\frac{N}{2}+N(\bar{p}-\frac{1}{2})]^2}{N}} && (12) \\
&= 1 - e^{-\frac{N^2(\bar{p}-\frac{1}{2})^2}{N}} && (13) \\
&= 1 - e^{-N\left(\frac{\sum_{i=1}^{N} p_i}{N} - \frac{1}{2}\right)^2}. && (14)
\end{aligned}$$

CCL 2024

**Algorithm 2** The annotation selection algorithm.

For each worker $w_j \in W$, maintain (1)a variable $\hat{\varphi}_j$ as the average $\mathrm{F}_1$ score of the selected annotations by $w_j$ so far, (2)a set $\boldsymbol{A}^j$ of selected annotations by $w_j$
Generate all possible annotations $\boldsymbol{A}_1^p$ on $s_1 \in S$, calculate $\mathrm{F}_1^{\mathrm{Exp}}(\boldsymbol{a}_{1k})$ for each $\boldsymbol{a}_{1k} \in \boldsymbol{A}_1^p$
For each $w \in W$, initialize $\hat{\varphi}_j$ with the $\mathrm{F}_1^{\mathrm{Exp}}(\boldsymbol{a}_{1k})$ closest to $\bar{\varphi}_j$, and append the $\boldsymbol{a}_{1k}$ to $\boldsymbol{A}^j$
**for all** $s_i \in S \backslash s_1$ **do**
  Generate all possible annotations $\boldsymbol{A}_i^p$ on $s_i \in S$, calculate $\mathrm{F}_1^{\mathrm{Exp}}(\boldsymbol{a}_{ik})$ for each $\boldsymbol{a}_{ik} \in \boldsymbol{A}_i^p$
  **for all** $w_j \in W$ **do**
    **if** $\hat{\varphi}_j > \bar{\varphi}_j$ **then**
      Update $\hat{\varphi}_j$ with the maximal $\mathrm{F}_1^{\mathrm{Exp}}(\boldsymbol{a}_{ik})$ less than $\bar{\varphi}_j$, and append $\boldsymbol{a}_{ik}$ to $\boldsymbol{A}^j$
    **else**
      Update $\hat{\varphi}_j$ with the minimal $\mathrm{F}_1^{\mathrm{Exp}}(\boldsymbol{a}_{ik})$ greater than $\bar{\varphi}_j$, and append $\boldsymbol{a}_{ik}$ to $\boldsymbol{A}^j$
    **end if**
  **end for**
**end for**

Through approximating $P_{\mathrm{MV}}$ by its lower bound in (14), we can see that the evaluation signal (represented by $P_{\mathrm{MV}}$) is an increasing function in each worker's capability $p_i$ and twice-differentiable. That is, $P_{\mathrm{MV}}$ is a generalized linear function, which satisfies Assumption 3 in (Garcelon et al., 2022). Therefore, one can follow the proof of Theorem 1 in (Garcelon et al., 2022) that the $\epsilon$-greedy algorithm yields a sub-linear regret with order $\tilde{O}(T^{2/3})$.

## C Case Study of Annotation Errors

Based on our statistical analysis of the Chinese OEI dataset, we find that 74.80% of annotations have different types of errors. And these annotation errors could be decomposed to three basic error types, namely Shifting, Expanding, and Shrinking (SES). In our data augmentation algorithm, we reversely used SES modifications and their combinations on the ground truth annotations to generate annotations with varying errors made by crowd workers. In this section, we provide a detailed characterization of human-made errors observed on annotated data with real cases to better motivate these modifications.

**Shifting** Some crowd annotation spans are as long as expert ones, but their positions are wrong. *Shifting* simulates this type of error. As depicted in Figure 5, both the expert span and the crowd span are three words long and of negative polarity. The difference is that the crowd span is shifted to the left by 2 words compared with the expert span. This type of error can be generated with *Shifting* on the expert annotations.

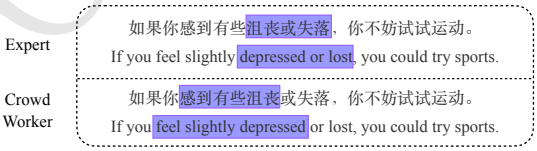



Figure 5: A case in which the crowd worker annotates a span with correct length and polarity but incorrect position.

**Expanding** *Expanding* is used to generate longer (than expert span) error spans. It might be intuitive that annotators barely make errors such as expanding to a very long span. However, in the case illustrated in Figure 6, the expert annotates five short spans separated by commas, while the crowd worker uses a very long span that covers the whole sentence, which is obviously not accurate. To simulate such human-made errors, we can expand an expert span to cover unnecessary words. Statistically, 4.03% of annotation errors are very long spans with more than 15 Chinese characters. So we do not set an upper bound of span length in *Expanding*.

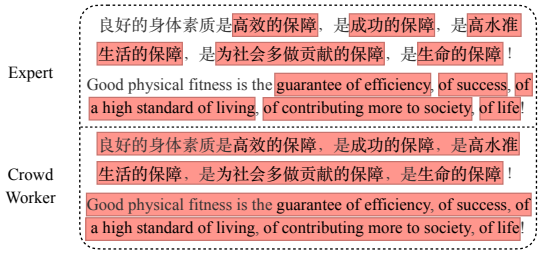



Figure 6: A case in which the crowd worker uses a very long span that covers the whole sentence.

**Shrinking** *Shrinking* is useful since crowd workers often ignore some words when annotating. As shown in Figure 7, the crowd worker failed to find all words expressing positive opinions.

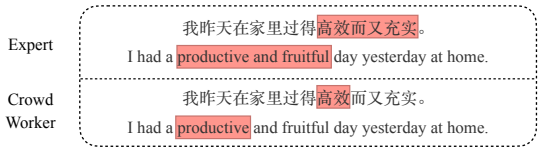



Figure 7: A case in which the crowd worker does not annotate all words with polarity.

Sometimes crowd workers ignore a whole span in expert annotations. This is why we set the lower bound of span length to zero in *Shrinking*, which means we can shrink a span into no span.

These three types of errors may occur separately or combined in real crowd annotations. Such that an error could be both shifted and shrunk. This is why we use the combination of these three types of modifications to simulate human-made errors in our data augmentation algorithm.

## D A Running Example of Data Augmentation

We here provide a running example to illustrate how an annotation for a certain worker on a certain sentence is generated with our proposed augmentation method. Suppose we have an English sentence:

> Although he looked very depressed yesterday, he has already become much more cheerful now.

And an expert annotation:

> Although he looked **[NEGATIVE: very depressed]** yesterday, he has already become **[POSITIVE: much more cheerful]** now.

If the crowd worker Sam has an annotation on this sentence in the original dataset, we use it directly in the augmented dataset. Otherwise, we generate an annotation for Sam with our data augmentation method.

When generating annotation for Sam, we follow the steps below:

1. For each span in the expert annotation, we apply the Shifting, Expanding, and Shrinking (SES) modifications on it. After this step, we have several lists of annotation, each list contain annotations with only one modified span:
   - List 1, modifications of the first span, containing $N_1$ annotations:
     - Although he looked **[NEGATIVE: very depressed]** yesterday, he has already become much more cheerful now. *# Unmodified, span-level proportional F1 = 1.0*
     - Although he looked very **[NEGATIVE: depressed yesterday]**, he has already become much more cheerful now. *# Shifting, F1 = 0.5*
     - Although he looked very depressed **[NEGATIVE: yesterday ,]** he has already become much more cheerful now. *# Shifting, F1 = 0*

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1239-1256, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

... *# Other Shifting modifications*
– Although he **[NEGATIVE: looked very depressed]** yesterday, he has already become much more cheerful now. *# Expanding, F1 = 1.0*
... *# Other Expanding modifications*
– Although he looked very **[NEGATIVE: depressed]** yesterday, he has already become much more cheerful now. *# Shrinking, F1 = 0.5*
... *# Other Shrinking modifications*

- List 2, modifications of the second span, containing $N_2$ annotations:
  – Although he looked very depressed yesterday, he has already become **[POSITIVE: much more cheerful]** now. *# Unmodified, span-level proportional F1 = 1.0*
  – Although he looked very depressed yesterday, he has already become much **[POSITIVE: more cheerful now]**. *# Shifting, F1 = 0.6667*
  – Although he looked very depressed yesterday, he has already become much more **[POSITIVE: cheerful now .]** *# Shifting, F1 = 0.3334*
  ... *# Other Shifting modifications*
  – Although he looked very depressed yesterday, he has already **[POSITIVE: become much more cheerful]** now. *# Expanding, F1 = 1.0*
  ... *# Other Expanding modifications*
  – Although he looked very depressed yesterday, he has already become much **[POSITIVE: more cheerful]** now. *# Shrinking, F1 = 0.6667*
  ... *# Other Shrinking modifications*

2. We choose one annotation from each list, and combine them to generate an annotation with 2 spans. This is done for all combinations of the annotations in the two lists. Note that if the two spans overlay with each other, we merge them into one span. After step 2, we have one list of annotations:

   Combined List, containing less than or equal to $N_1 \times N_2$ annotations:
   - Although he looked **[NEGATIVE: very depressed]** yesterday, he has already become **[POSITIVE: much more cheerful]** now. *# span-level proportional F1 = 1.0*
   - Although he looked very **[NEGATIVE: depressed yesterday]** , he has already become **[POSITIVE: much more cheerful]** now. *# span-level proportional F1 = 0.75*
   - Although he looked very depressed **[NEGATIVE: yesterday ,]** he has already become **[POSITIVE: much more cheerful]** now. *# span-level proportional F1 = 0.5*

   ... *# Other combinations with F1 ranging from 0 to 1.0*

3. We choose one annotation from the combined list as Sam's annotation on this sentence, according to the following procedure:
   (a) Sam has an average F1 score $F_{\text{ori}} = 0.57$ on the original (real) dataset.
   (b) We have already got 10 annotations for Sam in the augmented dataset, which has an average F1 score $F_{\text{aug\_10}} = 0.54$.
   (c) We are choosing annotation on the 11th sentence for Sam.
   (d) We firstly select two annotations with the closest F1 scores to $F_{\text{ori}}$ from the combined list, one higher than $F_{\text{ori}}$, and one lower than $F_{\text{ori}}$, as candidate annotations. In this case, the two annotations could have F1 scores of 0.58 and 0.52 respectively.
   (e) If $F_{\text{aug\_10}} > F_{\text{ori}}$, we choose the annotation with the lower F1 score (0.52) as Sam's annotation on this sentence. Otherwise, we choose the annotation with the higher F1 score (0.58). This is to ensure that the average F1 score of Sam's annotations in the whole augmented dataset, $F_{\text{aug}}$, is as close to $F_{\text{ori}}$ as possible, which reflects Sam's reliability (i.e., performance). In this case, we choose the annotation with F1 score of 0.58.

By generating the missing annotations in the original dataset with the method above, we could have an augmented dataset.

CCL 2024

## E Explanation of Worker Selection on Building New Datasets

When creating new datasets, we expect to have a few (e.g. five) experts and a relatively large group of (e.g. a hundred) crowd workers available for annotation.

At each time step, we select a group of (e.g. 20) crowd workers, and request them to annotate a few (e.g. 5) sentences, resulting in 4 crowd annotations on each sentence. Now we calculate the agreement of the annotations on each sentence, if the agreement is high (e.g. greater than 0.4), we use the MV aggregation of the crowd annotations as the ground truth, and calculate the F1 scores of each worker's annotation. Otherwise, we ask an expert to give an annotation on the sentence, and calculate the F1 score of each worker on the expert annotation. Note that the expert annotates only when the agreement is low. After this time step, we have crowd annotations on the sentences and their F-scores, which can be used to update the average score of each worker. This procedure is repeated until we have enough annotations on every sentence.

In other words, the Expert+MV approach works with both crowd workers and experts available (e.g. on an online system) when building datasets. The Expert+MV is an iterative approach in which the expert annotates when needed. And it saves the cost of expert annotations by using the MV aggregation of crowd annotations as the ground truth when possible. Our experiment results show that the Expert+MV approach can save 47.19% of the cost of expert annotations on the Chinese OEI dataset, and 65.97% on the CoNLL'03 dataset respectively.

However, even in the case that no expert is available, which means that Expert+MV falls back to MV, we can still observe that the MV approach outperforms the Random baseline (which is an equivalent of normal crowdsourcing procedure which assigns an equal amount of sentences to each worker randomly) by a large gap. In this case, the MV approach saves 100% of expert annotation cost, but still produced crowd annotation with good quality. Please refer to Table 3 and Table 4 for more detailed results.

CCL 2024

| Worker ID | Ori. $F_1$ | Rnd. Gen. $F_1$ | SES Only $F_1$ | SES +Alg.2 $F_1$ | Worker ID | Ori. $F_1$ | Rnd. Gen. $F_1$ | SES Only $F_1$ | SES +Alg.2 $F_1$ |
|---|---|---|---|---|---|---|---|---|---|
| 25 | 62.90 | 60.07 | 69.59 | 62.89 | 37 | 37.15 | 96.10 | 26.79 | 37.16 |
| 32 | 60.87 | 41.37 | 68.79 | 60.87 | 13 | 36.19 | 31.62 | 25.14 | 36.20 |
| 42 | 53.88 | 4.37 | 66.57 | 53.88 | 20 | 36.11 | 71.44 | 25.02 | 36.12 |
| 5 | 52.07 | 50.74 | 60.76 | 52.06 | 64 | 35.97 | 65.66 | 25.39 | 35.97 |
| 55 | 50.70 | 30.24 | 61.13 | 50.70 | 63 | 35.22 | 75.40 | 24.73 | 35.22 |
| 2 | 50.53 | 91.99 | 60.92 | 50.53 | 6 | 35.15 | 65.74 | 25.00 | 35.16 |
| 52 | 50.08 | 41.93 | 60.91 | 50.08 | 10 | 34.63 | 51.28 | 25.08 | 34.64 |
| 17 | 49.82 | 43.73 | 35.82 | 49.82 | 66 | 33.75 | 60.98 | 24.99 | 33.75 |
| 57 | 49.25 | 13.17 | 35.59 | 49.25 | 53 | 32.90 | 27.51 | 24.78 | 32.89 |
| 11 | 49.04 | 53.71 | 35.19 | 49.03 | 4 | 32.72 | 8.40 | 24.77 | 32.72 |
| 26 | 48.89 | 5.17 | 35.59 | 48.82 | 21 | 32.19 | 73.47 | 24.78 | 32.19 |
| 36 | 48.71 | 15.53 | 35.27 | 48.70 | 62 | 32.16 | 48.71 | 24.89 | 32.16 |
| 46 | 48.67 | 44.84 | 35.19 | 48.67 | 1 | 32.10 | 34.42 | 24.96 | 32.10 |
| 29 | 48.60 | 95.39 | 35.21 | 48.60 | 41 | 31.94 | 77.55 | 24.88 | 31.93 |
| 35 | 47.07 | 23.64 | 35.34 | 47.07 | 51 | 31.78 | 68.07 | 24.85 | 31.78 |
| 49 | 46.80 | 60.30 | 35.27 | 46.80 | 31 | 31.61 | 29.44 | 24.59 | 31.61 |
| 54 | 45.63 | 18.74 | 34.45 | 45.64 | 8 | 31.05 | 28.55 | 24.76 | 31.05 |
| 14 | 45.13 | 60.99 | 34.54 | 45.13 | 67 | 30.91 | 95.51 | 24.22 | 30.91 |
| 43 | 44.93 | 34.91 | 33.72 | 44.93 | 58 | 30.70 | 21.64 | 23.96 | 30.70 |
| 7 | 44.37 | 23.89 | 33.50 | 44.37 | 65 | 30.61 | 4.51 | 24.17 | 30.60 |
| 59 | 44.36 | 72.37 | 33.61 | 44.37 | 38 | 30.47 | 4.82 | 24.11 | 30.47 |
| 23 | 43.38 | 4.85 | 33.58 | 43.38 | 28 | 29.86 | 2.63 | 24.00 | 29.86 |
| 56 | 43.37 | 41.96 | 33.31 | 43.37 | 45 | 29.38 | 36.13 | 24.15 | 29.38 |
| 0 | 41.60 | 66.81 | 28.19 | 41.61 | 30 | 28.70 | 61.16 | 21.88 | 28.71 |
| 18 | 41.40 | 31.53 | 28.56 | 41.40 | 15 | 25.73 | 38.92 | 21.40 | 25.73 |
| 16 | 41.31 | 57.13 | 28.03 | 41.31 | 19 | 24.69 | 4.39 | 21.31 | 24.70 |
| 22 | 41.05 | 85.83 | 28.21 | 41.06 | 44 | 23.42 | 7.15 | 21.08 | 23.42 |
| 47 | 40.78 | 82.33 | 27.91 | 40.78 | 9 | 22.88 | 96.22 | 21.22 | 22.89 |
| 61 | 40.22 | 12.20 | 28.44 | 40.22 | 33 | 22.36 | 29.89 | 19.50 | 22.36 |
| 40 | 40.01 | 84.98 | 28.38 | 40.02 | 39 | 20.69 | 57.73 | 19.26 | 20.69 |
| 50 | 39.35 | 56.04 | 28.64 | 39.35 | 69 | 20.39 | 63.02 | 19.26 | 20.40 |
| 27 | 38.77 | 34.07 | 27.87 | 38.77 | 3 | 17.12 | 28.70 | 18.66 | 17.13 |
| 48 | 38.35 | 23.77 | 27.57 | 38.35 | 24 | 16.96 | 42.73 | 18.68 | 16.98 |
| 34 | 38.29 | 5.69 | 28.08 | 38.30 | 68 | 14.53 | 13.63 | 7.69 | 14.53 |
| 12 | 37.96 | 85.14 | 27.44 | 37.96 | 60 | 13.66 | 22.69 | 8.15 | 13.66 |

Table 7: Comparisons between different data augmentation methods on the span-level exact $F_1$ score of every crowd worker. **Ori.** stands for the original score in real datasets before any augmentation. **Rnd. Gen.** is a naive augmentation method with random generated annotations. **SES Only** indicates the *shifting*, *shrinking*, and *expanding* method we proposed. **SES + Alg.2** means SES with Algorithm 2 which is our final method.

ccl 2024

# DLUE: Benchmarking Document Language Understanding

**Ruoxi Xu**[1,2], **Hongyu Lin**[2]*, **Xinyan Guan**[1,2], **Yingfei Sun**[1]*, **Le Sun**[2,3]
[1]University of Chinese Academy of Sciences, Beijing, China
[2]Chinese Information Processing Laboratory [3]State Key Laboratory of Computer Science
Institute of Software, Chinese Academy of Sciences, Beijing, China
{ruoxi2021,hongyu,guanxinyan2022,sunle}@iscas.ac.cn
yfsun@ucas.ac.cn

## Abstract

Understanding documents is central to many real-world tasks but remains a challenging topic. Unfortunately, there is no well-established consensus on how to comprehensively evaluate document understanding abilities, which significantly hinders the fair comparison and measuring the progress of the field. To benchmark document understanding researches, this paper summarizes four representative abilities, i.e., document classification, document structural analysis, document information extraction, and document transcription. Under the new evaluation framework, we propose **Document Language Understanding Evaluation – DLUE**, a new task suite which covers a wide-range of tasks in various forms, domains and document genres. We also systematically evaluate six well-established transformer models and representative LLMs on DLUE, and find that due to the lengthy content, complicated underlying structure and dispersed knowledge, document understanding is still far from being solved in complex real-world scenarios.

## 1 Introduction

Documents are basic units of the organization of natural language (Buckland, 1997). Understanding the structures and the semantics of documents is the foundation for understanding news articles (Kiesel et al., 2019), scientific papers (Dasigi et al., 2021), government reports (Huang et al., 2021b) , stories (Kočiskỳ et al., 2018), etc. Evaluating how a machine intelligence system can read, analyze and generate documents is an important part of evaluating its natural language abilities, which has long been a critical direction in NLP field.

While standard benchmarks like GLUE (Shanahan et al., 2016) and SuperGLUE (Wang et al., 2019) have become a critical part of NLP community, they primarily focused on short utterances like sentences or paragraphs. However, documents are much more than bag-of-sentences, which usually focus on the central theme (Benamara et al., 2017), with underlying structure bound by complex linguistic elements (Parsing, 2009) and dependent knowledge dispersed across the whole text (Huang et al., 2021a). Therefore, these benchmarks cannot be used to evaluate document understanding due to its unique challenges: *First, documents usually have lengthy content*, i.e., they usually are much longer than sentences/paragraphs thus it is difficult to process them due to the computational memory/runtime limitation of current NN models. *Second, documents have underlying structures*, which play a critical role during understanding. For example, mining arguments from a document needs to model both the local coherence between sentences and the global interactions between claims and arguments, which cannot be accomplished by only exploiting sentence- or paragraph-level information. *Third, the knowledge in a document is usually dispersed beyond sentences/paragraphs*, which makes it necessary to model and explore document-level context. For example, long-distance within-document coreference resolution needs to integrate global information to find all expressions that refer to the same entity distributed in full text.

Recently, an increasing number of researches have paid attention to the evaluation of document understanding. Yi Tay et al. (2021) proposes the Long Range Arena (LRA), which contains two syn-

*Corresponding Authors
©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1257-1269, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers

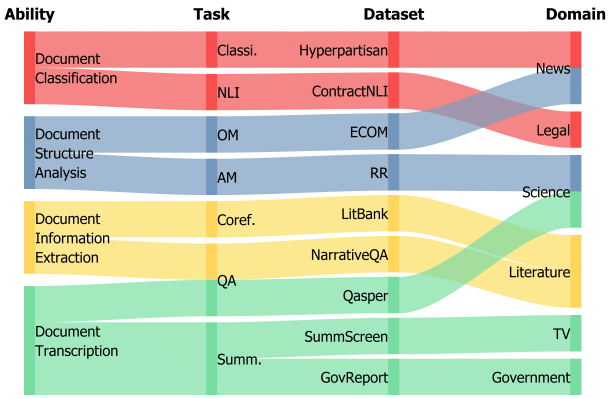



Figure 1: Overview of DLUE, which covers a wide-range of **tasks** and **datasets** in diverse **domains** to evaluate four representative document understanding **abilities**.

thetic text-related tasks and evaluates model quality of understanding multi-modal long contexts. Hudson and Moubayed (2022) proposes MuLD, which focuses on merged sequences over 10,000 tokens. SCROLLS (Shaham et al., 2022) is a benchmark that contains NLI, QA and summarization tasks and focuses on long language sequences, with ZeroSCROLLS (Shaham et al., 2023) as its extension for large language models. However, these benchmarks mainly focus on the lengthy content challenge, while ignoring other important challenges. As a result, almost all tasks in these benchmarks can be resolved via an retrieval-answering paradigm, i.e., retrieving a very limited number of sentences that contains critical information and then resolving the task. Furthermore, these benchmarks only cover limited tasks, which makes them unable to thoroughly evaluate the document understanding abilities of models.

To systematically evaluate document language understanding abilities, this paper proposes **Document Language Understanding Evaluation – DLUE**, a new task suite which covers a wide-range of tasks in various forms, different domains and document genres. Figure 1 shows the overview of DLUE. Specifically, we summarize 4 kinds of document understanding abilities, including 1) Document Classification, which evaluates whether a model can understand the overall semantics of a document, e.g., its topic (Zhang et al., 2015) and standpoint (Kiesel et al., 2019); 2) Document Structural Analysis, which evaluates whether a model can analyze and leverage the underlying structure of a document, e.g., its discourse structure (Zeldes, 2017) and argument structure (Cheng et al., 2020); 3) Document Information Extraction, which evaluates whether a model can recognize and aggregate associated information spanning cross whole document, e.g., long-distance within-document coreference (Bamman et al., 2020); 4) Document Transcription, which evaluates whether a model can capture and transcript important information of a document, e.g., summarization (Huang et al., 2021b; Chen et al., 2022) and abstractive QA (Dasigi et al., 2021). Then we collect 9 datasets and align them with the above abilities. Datasets of the same group are converted to a unified format. In this way, DLUE provides a comprehensive benchmark for document language understanding, which enables the research community to fairly compare and measure the progress of this field.

To better understand the challenges of document language understanding and analyze the performance of current approaches, we experiment on DLUE using several advanced document understanding models, including 1) Memory-based approaches, which includes XLNet (Yang et al., 2019); 2) Pattern-based approaches, which includes Longformer (Beltagy et al., 2020), BigBird (Zaheer et al., 2020), Sparse transformer (Child et al., 2019); 3) Low-rank/Kernel-based approaches, which includes Linformer (Wang et al., 2020) and Performer (Choromanski et al., 2020); 4) Large language models, which includes GPT-4, GPT-3.5-turbo[0], Vicuna-7B (Chiang et al., 2023), LLaMA-2-7B-chat (Touvron et al., 2023), Mistral-7B-instruct (Jiang et al., 2023) and ChatGLM3 (Du et al., 2022). Results show that document understanding is still far from being solved due to lengthy content, complex structure and dispersed knowledge, and currently there is no neural architecture that dominates all tasks, raising needs for a universal document

[0]https://openai.com/blog/openai-api

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1257-1269, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

understanding architecture.

Generally, the contributions[1] of this paper are:

- We summarize 4 representative abilities for lengthy, structural and global document understanding, including document classification, document structural analysis, document information extraction and document transcription.
- We propose a comprehensive benchmark for document language understanding – DLUE, which is built on established annotated datasets and selected to cover a diverse range of text genres, dataset sizes, and degrees of difficulty.
- We evaluate current state-of-the-art document understanding models on DLUE, which provides a novel reference to assess current models' abilities and properties when they handle different kinds of document understanding tasks.

## 2 Background

**NLP Benchmarks** The development of natural language understanding (NLU) evaluation benchmarks has helped drive the progress of pretraining and transfer learning in NLP. Benchmarks proposed in the early stage mostly aim at general tasks, such as SentEval (Conneau and Kiela, 2018) for universal sentence representations, DecaNLP (McCann et al., 2018) for ten diversified NLP tasks cast as a general question-answering format, GLUE (Wang et al., 2018) for NLU in English and SuperGLUE (Wang et al., 2019) as a harder counterpart of GLUE. Besides, benchmarks for more specific tasks have also been proposed, such as DialoGLUE (Mehri et al., 2020) for task-oriented dialogue, DiscoEval (Chen et al., 2019) for discourse-aware sentence representations, GLUECoS (Khanuja et al., 2020) for code-switched NLP, KILT (Petroni et al., 2020) for knowledge-intensive language tasks, and etc.

Above benchmarks mostly focus on sentences or short texts. However, documents are also very common for complex tasks or real-world textual data. Single-task benchmarks mostly use summarization (Cohan et al., 2018a) or QA tasks (Dasigi et al., 2021), but struggle to fully assess document modeling due to the single nature of the task and data distribution. There are also some multi-task benchmarks for document understanding, such as the Long Range Arena (LRA) (Tay et al., 2020a), SCROLLS (Shaham et al., 2022), ZeroSCROLLS (Shaham et al., 2023), MuLD (Hudson and Moubayed, 2022) and LOT (Guan et al., 2022). Most long inputs of LRA and MuLD are either automatically generated or artificially lengthened. Tasks in SCROLLS also tend to focus on a few sentences or paragraphs, which can be solved by retrieval-based or chunk-based approaches. LOT only focuses on Chinese text. In this paper, distinct from existing benchmarks focusing on long sequences instead of documents, we focus on challenges posed by document understanding, including lengthy content, complicated underlying structure and dispersed knowledge.

**Document Understanding Models** There have been numerous attempts to improve both memory footprint and computational cost of transformers, thus allowing use of longer inputs. A natural way is to connect blocks via recurrence, such as XLNet (Yang et al., 2019). Another way of tackling the high complexity of full attention is to sparsify attention matrix. Longformer (Beltagy et al., 2020), BigBird (Zaheer et al., 2020) and Sparse transformer (Child et al., 2019) simply sparsify the attention matrix by limiting the field of view to fixed, predefined patterns such as local windows and block patterns of fixed strides. Reformer (Kitaev et al., 2020) uses learnable ones, an extension to fixed, pre-determined pattern. Besides, low-rank approximations or kernelization of the self-attention matrix can be used as well to reduce the complexity. Linformer (Wang et al., 2020) and Performer (Choromanski et al., 2020) are representative low-rank/kernel-based transformers. In this paper, we conduct experiments on the above three document understanding architectures to explore challenges posed by document understanding.

## 3 DLUE: Document Language Understanding Evaluation Benchmark

The section describes the DLUE benchmark, used to evaluate 4 representative document understanding abilities. Specifically, DLUE is centered on 9 English document understanding datasets, which cover

[1] https://github.com/RossiXu/DLUEvaluation.git

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1257-1269, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

| Corpus | Task | Domain | Metric | Avg #Words Input | Avg #Words Output | #Examples |
|---|---|---|---|---|---|---|
| | | | Classification | | | |
| Hyperpartisan | Classifi. | News | acc. | 588 | 1 | 1273 |
| ContractNLI | NLI | Legal | acc. | 1708 | 1 | 10319 |
| | | | Structure Analysis | | | |
| ECOM | OP | News | $F_1$ | 488 | 20 | 2000 |
| RR | AM | Science | $F_1$ | 793 | 47 | 4764 |
| | | | Extraction | | | |
| LitBank | Coref. | Literature | $F_1$ | 2115 | 7.4 | 7214 |
| NarrativeQA | QA | Literature | $F_1$ | 51790 | 4.6 | 71187 |
| | | | Transcription | | | |
| GovReport | Summ. | Government | ROUGE | 7897 | 492.7 | 19402 |
| SummScreen | Summ. | TV | ROUGE | 5639 | 100.0 | 4348 |
| Qasper | QA | Science | $F_1$ | 3671 | 11.5 | 5692 |

Table 1: Task descriptions and statistics of DLUE.

a wide-range of tasks in various forms, different domains and document genres. Table 1 provides an overview of the datasets included in the benchmark. In the following, we describe the details of DLUE.

### 3.1 Overview

As described above, a document understanding system should resolve the lengthy content, complicated underlying structure, and dispersed knowledge challenges. To effectively evaluate the above abilities and challenges, DLUE selects datasets based on these criteria: First, documents should have lengthy content. Second, tasks require leveraging dispersed document knowledge; hence, document-level datasets aren't preferred if most of them can be tackled through chunk- or retrieval-based methods. Third, documents should be natural, like literature, scientific articles, or government reports. Synthesized documents lack structure information and relation links among different sections. Fourth, tasks should exceed current system capabilities but remain solvable by most educated English speakers. Based on the above desiderata and with the permission of licenses, we collect as diverse datasets as possible to increase the coverage on capabilities. The overview of DLUE is shown in Figure 1, and their statistics are shown in Table 1. In the following, we describe all datasets according to the their target ability.

### 3.2 Document Classification

A document usually narrow focus on a single central theme (Benamara et al., 2017). We aim to evaluate document classification ability, specifically the ability to understand the overall semantics of documents in this section. To do this, we select two datasets that rely on full-text to make judgements and reformulate every dataset as document classification tasks. Specifically, given single sequence $s$ or sequence pairs $(s_1, s_2)$, the goal is to classify the input into a single label $l$.

**Hyperpartisan** (Kiesel et al., 2019) is a document classification dataset for classifying news based on left-wing or right-wing standpoints. A few words or sentences aren't sufficient to gauge the political leaning, which are toned by the full text. This task provides two datasets, one labeled manually and the other labeled semi-automatically via distant supervision at the publisher level. We use the first one to pursue higher evaluation accuracy and keep the same train/test split as the original work.

**ContractNLI** (Koreeda and Manning, 2021) is a natural language inference dataset in the legal domain, with non-disclosure agreements (NDAs) as premises and legal statements as hypothesizes. NDAs are collected from Internet search engines and Electronic Data Gathering, Analysis, and Retrieval system. To correctly predict whether the hypothesis is entailed, neutral, or contradictory from the contract, we need to refer to non-continuous sentences across the contract with hundreds of tokens. The dataset contains 607 contracts and 17 hypothesizes, which we combine to produce 10319 instances.

### 3.3 Document Structure Analysis

A document is composed of structured groups of sentences, paragraphs and sections. Analyzing document structure can be very useful in indexing and organizing the information contained in the document.

CCL 2024

Tasks in this section aim to evaluate document structure analysis ability, specifically the ability to capture and leverage structure information. We select three datasets as follows and reformulate every dataset as sentence-level sequence labeling tasks. Specifically, given a document $d = \{s_1, s_2, ..., s_n\}$, the goal is to output a tag sequence $t = \{t_1, t_2, ..., t_n\}$ for sentences.

**ECOM** (Xu et al., 2022) is an event-centric opinion mining corpus in which a system takes in an event descriptor and related news articles to extract event-related opinion segments from articles. An opinion segment is composed of continuous sentences targeting at the same argument. We select the dataset to evaluate the ability to utilize local structure information, which is important for identifying opinion boundaries unambiguously.

**RR** (Cheng et al., 2020) is an argument mining corpus for extracting arguments from reviews and rebuttals, sourced from ICLR spanning 2013 to 2020 (except 2015). Peer reviews and rebuttals on scientific works are a data source of rich structures and long passages. It's considered suitable because the original paper's experiments highlight the significance of internal structure information for this task.

### 3.4 Document Information Extraction

Dependent knowledge in a document is usually dispersed across the full text, which plays an important role in the transmission of the intended meaning. Tasks in this section aim to evaluate document information extraction ability, specifically the ability to identify long-distance related mentions and relations. We select two datasets as follows and reformulate every dataset as multi-answer question answering tasks. Specifically, given a document $d$ and a question $q$, the goal is to extract correct answer spans $a = \{a_1, a_2, ..., a_n\}$ from $d$ for $q$.

**LitBank** (Bamman et al., 2020) is a coreference resolution dataset on English literature. Its documents are several times longer than those in other benchmarks (e.g. 463.7 tokens for OntoNotes), making them rich in long-distance within-document coreference. For each coreference link, we transform the sentence of one mention into a question, take all mentions as answers, resulting in 7214 question-answer pairs.

**NarrativeQA** (Kočiskỳ et al., 2018) is a reading comprehension dataset on books and movie scripts. Its questions are written based on summaries, requiring readers to grasp long-distance associated information across several parts or a larger span of the context document for comprehension and answering.

### 3.5 Document Transcription

Tasks in this section aim to evaluate document transcription ability, specifically the ability to capture and transcript key information of documents. We select three datasets that need to contextualize across different sections and reformulate every dataset as sequence-to-sequence tasks. Specifically, given a sequence $s$, the goal is to output a concise and fluent new sequence $s_N$.

**GOVREPORT** (Huang et al., 2021b) is a summarization dataset of long reports on national policy issues and paired expert-written summaries. Its documents and summaries are significantly longer than prior datasets, such as 1.5 times longer than Arxiv (Cohan et al., 2018b). Moreover, new salient bigrams are steadily added as more content is consumed, suggesting information is spread throughout documents in the dataset.

**SummScreen** (Chen et al., 2022) is a summarization dataset of TV series transcripts paired with human written recaps. Unlike official documentation like GOVREPORT (Huang et al., 2021b), its language expression is informal and structure is more unclear. Whole documents need to be combined to understand plots often expressed indirectly in character dialogues and scattered throughout the transcript.

**Qasper** (Dasigi et al., 2021) is a QA dataset in the research domain focusing on entire papers, in which both questions and answers are handed-written by NLP practitioners. Over half of questions require multiple paragraphs as evidence to answer. We prepend the query to the document, using two newlines as a natural separator to construct the input sequence.

CCL 2024

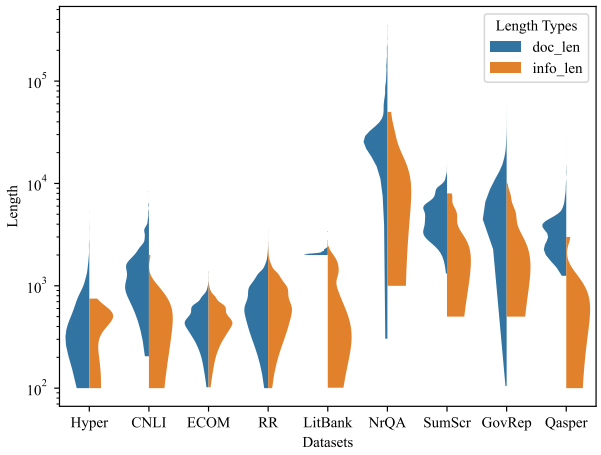



Figure 2: The distribution of document lengths and information spread in DLUE datasets. The spread of information measured by the standard deviation of the position of each bigram's first occurrence in the input document like (Shaham et al., 2022).

### 3.6 Data Analysis

Datasets in DLUE are analyzed, including documents' length, and distribution of information. In Figure 2, we can observe that the benchmark consist of documents with varying lengths and essential information required for document understanding is distributed throughout the entire text.

## 4 Experiments and Analysis

### 4.1 Benchmarking Architectures

This section describes models and architectures we evaluate on DLUE. Specifically, we compare three efficient transformer architectures following the general taxonomy (Tay et al., 2020b) including memory-based models (Yang et al., 2019), pattern-based models (Zaheer et al., 2020; Beltagy et al., 2020; Child et al., 2019), low-rank/Kernel-based models (Wang et al., 2020; Wang et al., 2020) and also large language models.

For all kinds of task formulations described in Section 3, we implement unified model architectures. For document classification tasks, we use the special classification token ([CLS]) for prediction. The document structure analysis tasks are reformulated into sentence-level sequence labeling tasks. We use the classical Transformer-CRF architecture as in named entity recognition (Devlin et al., 2019). The document information extraction tasks are reformulated into multi-span question answering tasks. Following Hu et al. (2019), we expand traditional MRC architecture by adding a span number prediction and search component. For transcription tasks, we use the basic encoder-decoder architecture (Vaswani et al., 2017).

### 4.2 Implementations

Our models are implemented by PyTorch framework[2][3]. For transformers with public pretrained models, we use the base version, including XLNet-base, Longformer-base, BigBird-base. The learning rate is 1e-5 for pretrained models and 1e-3 for classifier heads. For other models, we follow the same setup as Long range arena (Tay et al., 2020a), a widely recognized benchmark for efficient transformers to minimize the influence of hyper-parameter settings. These transformer models are parameterized by the same number of layers, heads and hidden dimensions, namely 6 layers, 8 heads, 512 hidden dimensions and d = 2048 for positional FFN layers. We use Adam with warmup. All models are trained for 10 epochs. Across datasets and models, we run three repetitions with different random seeds and report averaged scores.

[2]https://pytorch.org/

[3]In practice, these models often use a combination of the proposed approximate global attention and simple local attention.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1257-1269, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

| Model | #param | pretrain | Classification | | Structure Analysis | | Extraction | | Transcription | | | Avg | Inference Speed (steps per sec) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Hyper | CNLI | ECOM | RR | LitBank | NrQA | SummScr | GovRep | Qasper | | |
| | | | Vanilla Transformer | | | | | | | | | | |
| BERT | 110M | yes | 80.1 | 72.3 | 37.3 | 57.3 | - | 14.5 | - | - | - | - | - |
| | | | Memory-based | | | | | | | | | | |
| XLNet | 110M | yes | 81.4 | 80.2 | **39.1** | **74.0** | 78.1 | 15.2 | 18.9 | 22.7 | 24.2 | 41.1 | 0.95 |
| | | | Pattern-based | | | | | | | | | | |
| Longformer | 148M | yes | 83.8 | 71.6 | 37.9 | 72.9 | **79.1** | 18.3 | **20.9** | 25.7 | 26.4 | 42.1 | 1.7 |
| BigBird | 127M | yes | **85.9** | 82.8 | 37.0 | 71.1 | 77.8 | 18.2 | 20.6 | 27.3 | 26.2 | **42.7** | 1.6 |
| Sparse Trans. | 46M | no | 64.6 | 67.7 | 21.9 | 45.6 | 56.7 | 11.1 | 21.4 | 17.6 | 17.6 | 30.5 | 3.8 |
| | | | Low-rank/Kernel-based | | | | | | | | | | |
| Linformer | 45M | no | 67.1 | 65.5 | 22.6 | 44.3 | 63.8 | 12.4 | 18.9 | 25.8 | 17.5 | 30.5 | 6.4 |
| Performer | 44M | no | 67.9 | 69.5 | 18.6 | 48.6 | 51.6 | 10.1 | 20.1 | 15.6 | 21.5 | 30.9 | 6.7 |
| | | | Large Language Models | | | | | | | | | | |
| GPT-4 | - | yes | 79.4 | **88.0** | 6.9 | 7.5 | 74.3 | 5.1 | 18.0 | **31.1** | **53.5** | 40.4 | - |
| GPT-3.5-turbo | - | yes | 66.7 | 74.0 | 9.9 | 4.4 | 60.6 | **21.8** | 22.8 | 24.0 | 40.0 | 36.0 | - |
| Vicuna-7b | 7B | yes | 46.0 | 41.5 | 10.0 | 0.5 | 28.2 | 1.7 | 11.4 | 21.3 | 11.2 | 19.1 | - |
| Llama-2-7b-chat | 7B | yes | 40.9 | 41.5 | 7.9 | 1.6 | 24.9 | 2.3 | 11.9 | 17.3 | 6.7 | 17.2 | - |
| Mistral-7b-instruct | 7B | yes | 41.7 | 42.9 | 10.8 | 6.5 | 46.1 | 2.6 | 8.5 | 25.2 | 8.0 | 21.4 | - |
| ChatGLM3-6b | 6B | yes | 57.4 | 58.1 | 13.2 | 6.1 | 22.2 | 3.0 | 11.1 | 28.4 | 10.6 | 23.3 | - |

Table 2: Overall experimental results on DLUE. Best model is in boldface. "-" denotes the model can't handle this task. LLMs are tested in in a zero-shot setting to ensure a fair and unbiased comparison.

### 4.3 Overall Results

Table 2 shows the overall results on DLUE. From this table, we can see that:

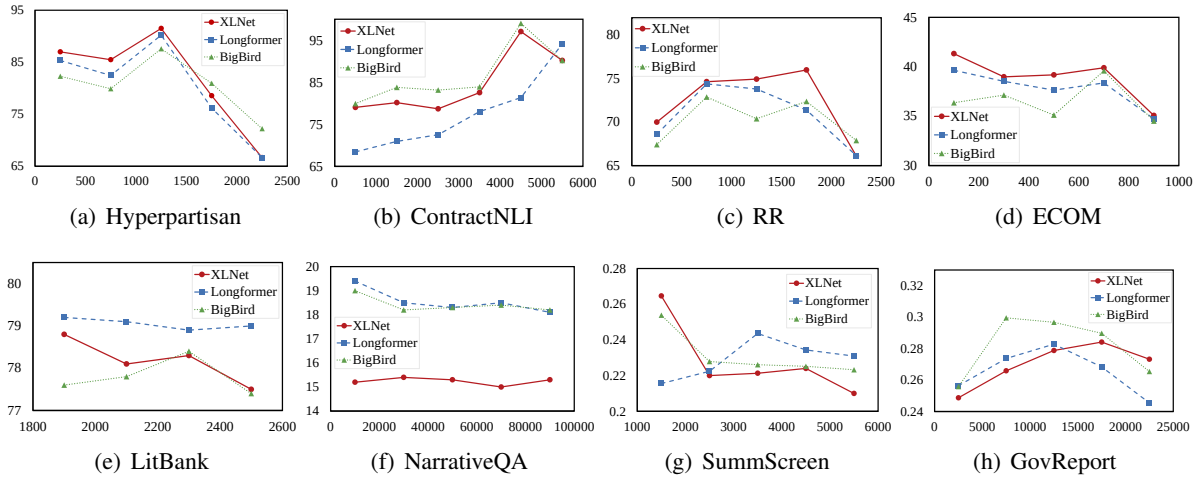
**1) Document understanding is far from being solved.** From Table 2, we can see that the best benchmark system can only achieve 42.7 average score. While it's difficult to establish an accurate human performance ceiling in DLUE, we can take some indicators to prove that the performance gap between human and models are huge. For example, human agreement on ECOM was measured at 80.8% F1 (Xu et al., 2022), much higher than our best baseline of 39.1% F1. Likewise, Dasigi et al. (2021) study a subset of Qasper that has multiple annotated answers, and find their overlap to be 60.9% F1, more than double our best baseline. This indicates that contemporary off-the-shelf models struggle with documents, challenging future work to make progress on DLUE.

**2) Different tasks have different advantageous architectures, raising a need for an universal document understanding architecture which can dominate all tasks in one architecture.** From Table 2, we can see that different model architectures seem to be good at processing different tasks. Specifically, the performance of XLNet ranks first on the structure analysis tasks, while Longformer and BigBird perform better on the other tasks. Linformer and Performer do well on document classification tasks. This shows that recurrence-based models may have advantages over hierarchically structured data and pattern-based models may be more effective on flat data. Contrary to the other tasks, fast low-rank/kernel-based models do better on document classification tasks. No architecture dominates all tasks, which indicates that more universal models are needed.

**3) Lengthy content is the critical, but not the only, challenge for document understanding.** From Table 2 and Table 1, we can see that models perform poorly with too long inputs, such as the 18.5 best $F_1$ score in NarrativeQA dataset with 51790 average input length. However, even for those structure analysis and extraction tasks where documents can be taken in completely by long-range transformer models, the model performances still fail to meet expectations. Obviously, there exist other challenges for document understanding apart from lengthy input, such as complex structures and dispersed knowledge.

**4) It is critical to take global context into consideration.** From Table 2, we can see that long-range transformers that can take in more contexts achieve a higher score than vanilla transformer in most datasets. This demonstrates longer contexts are necessary to understand documents. Document-level tasks can't be solved in the same way as short-text tasks.

CCL 2024




(a) Hyperpartisan (b) ContractNLI (c) RR (d) ECOM

(e) LitBank (f) NarrativeQA (g) SummScreen (h) GovReport

Figure 3: Performance ($y$ axis) on DLUE datasets with different document lengths ($x$ axis).

### 4.4 Computational Efficiency

The last column of Table 2 shows inference speeds of models. For a fair comparison, we use the standard test datasets of DLUE as testbed. Based on our implementation, the low-rank/kernel-based models are the fastest. The results are consistent with model complexity, which has a significant impact on inference speed. The low-rank/ kernel-based models decompose the $N \times N$ self-attention matrix to a lower-dimensional representation and thus usually have a $O(N)$ time complexity. Pattern-based models sparsify the attention matrix according to predefined or learnable patterns and the time complexity is usually between $O(N)$ and $O(N^2)$. Recurrence-based models connect multiple segments and blocks via recurrence and the representative XLNet has a $O(N^2)$ time complexity.

### 4.5 Effect Of Document Length

To investigate how the document length will impact the performance, we cluster the documents into buckets for each task according to their document lengths and run the evaluation on each bucket. The breakdown analysis is shown in Figure 3. We find that: 1) On the whole, understanding longer documents faces more challenges. We notice that the performances on most datasets decrease when document lengths increase, with ContractNLI dataset as an exception. This maybe because there exists label bias related to document lengths in ContractNLI datasets. We find that a longer contract tends to entail a hypothesis, with 34% probability for documents shorter than 1000 words and 76% probability for documents longer than 5000 words. 2) The performance of pattern-based models seems to be more stable when the document lengths increase. We can see that Longformer and BigBird obtain a greater advantage when documents get longer. We think there are two reasons. First, the global token mechanism in Longformer and BigBird could help models focus on important information and be less distracted by noise in long contexts. Second, the maximum input length of XLNet is smaller due to the segment-level recurrence mechanism. 3) The performance is relatively stable on datasets where document lengths far exceed input limits. Figure 3(f) shows performance on NarrativeQA dataset. When the document length exceeds 20,000 tokens, the result remains around 18 $F_1$ for Longformer, BigBird and 15 $F_1$ for XLNet. This indicates the ability of efficient transformers to understand long documents are limited.

### 4.6 Effect of dispersed knowledge Exploition

In this section, we aim to validate the importance of recognizing and aggregating dispersed knowledge for document understanding, with much room for current models to improve. We analyze from two perspectives: 1) effect of mention distance, which can be viewed as the measure of dispersion; 2) performance comparison between long-range transformers and short-text models without global information.

**Effect of Mention Distance** To quantify the impact of dispersed knowledge to document understanding, we analyze the performance of coreference resolution with different mention distances on LitBank.

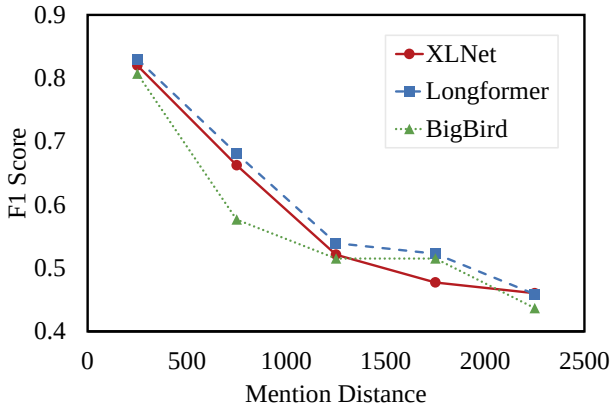



Figure 4: Performance on LitBank dataset with different mention distances. Mention distances can reflect the degree of knowledge dispersion in a document.

| Model | Hyperpartisan | Qasper |
|---|---|---|
| XLNet | 81.4 | 24.2 |
| Longformer | 83.8 | **26.4** |
| BigBird | **85.9** | 26.2 |
| CogLTX | 82.9 | 18.9 |
| ToBERT | 78.4 | 16.6 |

Table 3: Performance comparison between long-range transformers and short-text models.

In Figure 4, we can see performance of all models drops sharply with greater mention distances, indicating long-distance coreference is more challenging than within-sentence coreference. It's easy to understand because it puts forward higher requirements for ability to capture and aggregate dispersed information. We can also notice huge performance gap between short and long mention distances, implying there is still much room for further improvements in models' global information integration ability.

**Comparison with Short-text Models** To verify the importance of global information to document understanding, we compare the performance of long-range transformers with two existing short-text models, including CogLTX (Ding et al., 2020) and ToBERT (Pappagari et al., 2019). CogLTX jointly trains two BERT models to select key sentences from documents. ToBERT divides documents into smaller chunks and uses a transformer layer over BERT-based chunk representations. We select Hyperpartisan and Qasper datasets, whose tasks can be solved by CogLTX and ToBERT, and in which documents can be completely taken in by long-range transformers to eliminate interference caused by more contexts.

From Table 3, we can see that long-range transformers do have advantages over IR-based and chunking-based methods. Intuitively, the reason behind is that the performance of long-range transformers benefits from the contextual representation with a broader view of the document. These findings emphasize the need for future studies in document understanding to integrate global information. The results also indicate that DLUE effectively covers the assessment of the ability to recognize and aggregate dispersed knowledge across the whole text.

## 5 Conclusions

We propose a new benchmark DLUE that places the spot on documents and their lengthy content, complex underlying structure and dispersed knowledge challenges. DLUE covers diverse document-level tasks to evaluate four basic abilities required by document understanding, including document classification, document structure analysis, document information extraction and document transcription. Based on DLUE, we conduct an extensive side-by-side comparison of three document understanding architec-

CCL 2024

tures. Experiments demonstrate document understanding is far from being solved, and there exists a need for a universal architecture that can dominate all tasks.

## Limitations

DLUE now focuses on plain text documents, while the documents one encounter, e.g., scientific articles, company announcements, or even personal notes, may also contain multi-modal information and with non-sequential structure. In future work, we intend to integrate these multi-modal, complex structure information into our document understanding benchmark.

Besides, due to the huge cost of computing resources, we didn't pretrain models specialized for document understanding, but directly use the public pretrained versions or train from scratch. We believe an unified pretraining incorporating document-related tasks will enhance understanding performance.

## Ethics Statement

In consideration of ethical concerns, we provide the following detailed description:

1. We believe that this work is beneficial to develop universal document understanding architectures, which can help people quickly get information from business documents, legal statements and so on, saving time and money.

2. We standardize and put together ten datasets, which are all already publicly available under CC-BY-(NC-)SA-4.0 licenses[4]. For all the datasets, we have referenced the original work and encouraged DLUE users to do so.

3. All DLUE benchmark datasets have low ethical risks and do not expose any sensitive or personally identifiable information.

## Acknowledgements

We thank all reviewers for their valuable comments. Moreover, this work was supported by the National Natural Science Foundation of China under Grants no. 62122077, 62106251 and 62306303.

## References

David Bamman, Olivia Lewke, and Anya Mansoor. 2020. An annotated dataset of coreference in English literature. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 44–54, Marseille, France, May. European Language Resources Association.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Farah Benamara, Maite Taboada, and Yannick Mathieu. 2017. Evaluative language beyond bags of words: Linguistic insights and computational applications. *Computational Linguistics*, 43(1):201–264.

Michael K Buckland. 1997. What is a "document"? *Journal of the American society for information science*, 48(9):804–809.

Mingda Chen, Zewei Chu, and Kevin Gimpel. 2019. Evaluation benchmarks and learning criteria for discourse-aware sentence representations. *arXiv preprint arXiv:1909.00142*.

Mingda Chen, Zewei Chu, Sam Wiseman, and Kevin Gimpel. 2022. SummScreen: A dataset for abstractive screenplay summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8602–8615, Dublin, Ireland, May. Association for Computational Linguistics.

Liying Cheng, Lidong Bing, Qian Yu, Wei Lu, and Luo Si. 2020. Ape: argument pair extraction from peer review and rebuttal via multi-task learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7000–7011.

[4]https://creativecommons.org/licenses/by/4.0/

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. 2020. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018a. A discourse-aware attention model for abstractive summarization of long documents. *arXiv preprint arXiv:1804.05685*.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018b. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana, June. Association for Computational Linguistics.

Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. *arXiv preprint arXiv:1803.05449*.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4599–4610, Online, June. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Ming Ding, Chang Zhou, Hongxia Yang, and Jie Tang. 2020. Cogltx: Applying bert to long texts. *Advances in Neural Information Processing Systems*, 33:12792–12804.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Jian Guan, Zhuoer Feng, Yamei Chen, Ruilin He, Xiaoxi Mao, Changjie Fan, and Minlie Huang. 2022. Lot: A story-centric benchmark for evaluating chinese long text understanding and generation. *Transactions of the Association for Computational Linguistics*, 10:434–451.

Minghao Hu, Yuxing Peng, Zhen Huang, and Dongsheng Li. 2019. A multi-type multi-span network for reading comprehension that requires discrete reasoning. *arXiv preprint arXiv:1908.05514*.

Kung-Hsiang Huang, Sam Tang, and Nanyun Peng. 2021a. Document-level entity-based extraction as template generation. *arXiv preprint arXiv:2109.04901*.

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021b. Efficient attentions for long document summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1419–1436, Online, June. Association for Computational Linguistics.

G Thomas Hudson and Noura Al Moubayed. 2022. Muld: The multitask long document benchmark. *arXiv preprint arXiv:2202.07362*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. 2020. Gluecos: An evaluation benchmark for code-switched nlp. *arXiv preprint arXiv:2004.12376*.

Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. 2019. Semeval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839.

Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*.

Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Yuta Koreeda and Christopher D Manning. 2021. Contractnli: A dataset for document-level natural language inference for contracts. *arXiv preprint arXiv:2110.01799*.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*.

Shikib Mehri, Mihail Eric, and Dilek Hakkani-Tur. 2020. Dialoglue: A natural language understanding benchmark for task-oriented dialogue. *arXiv preprint arXiv:2009.13570*.

Raghavendra Pappagari, Piotr Zelasko, Jesús Villalba, Yishay Carmiel, and Najim Dehak. 2019. Hierarchical transformers for long document classification. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 838–844. IEEE.

Constituency Parsing. 2009. Speech and language processing.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, et al. 2020. Kilt: a benchmark for knowledge intensive language tasks. *arXiv preprint arXiv:2009.02252*.

Uri Shaham, Elad Segal, Maor Ivgi, Avia Efrat, Ori Yoran, Adi Haviv, Ankit Gupta, Wenhan Xiong, Mor Geva, Jonathan Berant, et al. 2022. Scrolls: Standardized comparison over long language sequences. *arXiv preprint arXiv:2201.03533*.

Uri Shaham, Maor Ivgi, Avia Efrat, Jonathan Berant, and Omer Levy. 2023. Zeroscrolls: A zero-shot benchmark for long text understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7977–7989.

Timothy Shanahan, Douglas Fisher, and Nancy Frey. 2016. The challenge of challenging text. *On developing readers: Readings from educational leadership (EL Essentials)*, 100.

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. 2020a. Long range arena: A benchmark for efficient transformers. *arXiv preprint arXiv:2011.04006*.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2020b. Efficient transformers: A survey. *ACM Computing Surveys (CSUR)*.

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. 2021. Long range arena : A benchmark for efficient transformers. In *International Conference on Learning Representations*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

CCL 2024

Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. 2020. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*.

Ruoxi Xu, Hongyu Lin, Meng Liao, Xianpei Han, Jin Xu, Wei Tan, Yingfei Sun, and Le Sun. 2022. Eco v1: Towards event-centric opinion mining. *arXiv preprint arXiv:2203.12264*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33:17283–17297.

Amir Zeldes. 2017. The gum corpus: Creating multilayer resources in the classroom. *Lang. Resour. Eval.*, 51(3):581–612, sep.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

CCL 2024

# Do Large Language Models Understand Conversational Implicature – A case study with a Chinese sitcom

**Shisen Yue  Siyuan Song  Xinyuan Cheng  Hai Hu***
School of Foreign Languages, Shanghai Jiao Tong University
shisenyue@gmail.com, {sjtusongsy2022, 0106cxy, hu.hai}@sjtu.edu.cn
*corresponding author

## Abstract

Understanding the non-literal meaning of an utterance is critical for large language models (LLMs) to become human-like social communicators. In this work, we introduce SwordsmanImp, the first Chinese multi-turn-dialogue-based dataset aimed at conversational implicature, sourced from dialogues in the Chinese sitcom *My Own Swordsman*. It includes 200 carefully handcrafted questions, all annotated on which Gricean maxims have been violated. We test eight close-source and open-source LLMs under two tasks: a multiple-choice question task and an implicature explanation task. Our results show that GPT-4 attains human-level accuracy (94%) on multiple-choice questions. CausalLM demonstrates a 78.5% accuracy following GPT-4. Other models, including GPT3.5 and several open-source models, demonstrate a lower accuracy ranging from 20% to 60% on multiple-choice questions. Human raters were asked to rate the explanation of the implicatures generated by LLMs on their reasonability, logic and fluency. While all models generate largely fluent and self-consistent text, their explanations score low on reasonability except for GPT-4, suggesting that most LLMs cannot produce satisfactory explanations of the implicatures in the conversation. Moreover, we find LLMs' performance does not vary significantly by Gricean maxims, suggesting that LLMs do not seem to process implicatures derived from different maxims differently. Our data and code are available at https://github.com/sjtu-compling/llm-pragmatics.

## 1 Introduction

The complexity of communication is largely epitomized by indirect, or non-literal utterances. A common instance is hinting at a busy schedule as a polite refusal to engage in an unwanted activity. How such implied meaning is understood in human communication has lone been a key subject of investigation in pragmatics research (Grice, 1975; Searle et al., 1980; Brown & Levinson, 1987; Wilson & Sperber, 2006).

Evaluating the pragmatic understanding ability of large language models (LLMs) has drawn considerable attention in recent years as LLMs show remarkable ability for language understanding. Recent studies have evaluated LLMs' pragmatic reasoning in multiple aspects, including scalar inference (Hu et al., 2023b), discourse connectives (Pandia et al., 2021), gradable adjectives (Lipkin et al., 2023) and conversational implicatures (Qiu et al., 2023; Kim et al., 2023; Ruis et al., 2022; Hu et al., 2023a; Zheng et al., 2021).

However, the above-mentioned evaluation are primarily in English, leaving a gap for pragmatic understanding in other languages. Moreover, previous conversational reasoning datasets involve few turns of dialogue, while our daily communication usually involves much more context and turn taking. In addition, previous studies mostly employ binary- or multiple-choice questions (Hu et al., 2023a; Ruis et al., 2022), which is inadequate in the era of LLM where it is possible to ask LLMs to generate explanatory text of the situation and directly evaluate its quality.

©2024 China National Conference on Computational Linguistics
Published under Creative Common Attribution 4.0 International License

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1270-1285, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

| Question | Choices | |
|---|---|---|
| **Dialogue:**<br>小郭：知道这意味着什么吗？以后再敢胡来，就不光是挨顿打的事了。<br>薇：烦死了，在家就是听爹娘罗嗦，好不容易溜出来，倒听你罗嗦。<br>小郭：什么？你是溜出来的？<br>薇：我说的那个溜，是溜冰的溜啊，现在京城都在下雪，满地都是冰，我是穿着冰刀，溜出来的，嘿嘿……<br>**English Translation:**<br>Xiaoguo: Do you know what this implies? If you dare to mess around again, it won't just be about getting a beating.<br>Wei: So annoying. At home, I have to listen to my parents nagging. Finally sneaking out, now I have to listen to you nagging.<br>Xiaoguo: What? You sneaked out?<br>Wei: When I said "sneak out,"[1] I meant ice skating. It's snowing in the capital now, the ground is covered in ice. I wore ice skates and skated out, hehe... | **Pragmatic** | 薇发现自己说溜出来说漏嘴了，就找补说是溜冰出来的。<br>Wei realized she accidentally slipped up by saying she sneaked out and tried to cover it up by saying she meant ice skating. |
| | **Literal** | 薇是从京城的家里溜冰出来的。<br>Wei sneaked out from her home in the capital to go ice skating. |
| | **Distractor#1** | 薇很喜欢溜冰。<br>Wei really enjoys ice skating. |
| | **Distractor#2** | 薇怕下雪冷，所以离开了京城。<br>Wei was afraid of the cold due to the snow, so she left the capital. |

**Maxim Check**

| Maxim | *Sub-maxims* |
|---|---|
| Quality | ✗ Do not say what you believe is false.<br>✗ Do not say that for which you lack adequate evidence |
| Quantity | ✓ Make your contribution as informative as is required.<br>✗ Do not make your contribution more informative than is required. |
| Relation | ✓ Be relevant |
| Manner | ✓ Avoid obscurity of expression.<br>✓ Avoid ambiguity<br>✗ Be brief<br>✓ Be orderly |

Figure 1: An example entry in our dataset. The tick (✓) and fork (✗) denotes if the target sentence, marked in red, comply with or violate the sub-maxim. This entry belongs to the classes of Quality, Quantity and Manner. "sneak out" and "skate" translate to the same Chinese character "溜".

To bridge the aforementioned gaps, we present SwordsmanImp, the first Chinese multi-turn-dialogue-based dataset aimed at conversational implicature. It consists of 200 dialogues carefully excerpted by linguistic experts from the Chinese sitcom *My Own Swordsman* (武林外传). Figure 1 shows an example dialogue in the dataset. Each dialogue contains one sentence that carries a non-literal meaning. We provide four well-organized interpretations of this sentence, including a pragmatic meaning, a literal understanding, and two incorrect inferences that involve related information (distractors).

From a linguistic perspective, many pragmatic inferences, especially conversational implicature, arise because the speaker has violated one or more maxims in the cooperative principle (Grice, 1989). For example: when asked "Is John in the office?", Sam replied "It's Saturday, you know". This violates the Maxim of Relation since the reply is not directly related to the question, which then gives rise to an implicature: "John never works on weekends, so he is not in the office". Meanwhile, it has been shown that the difficulty in detecting conversational implicature hinges crucially on which maxims have been violated (Engelhardt et al., 2006; Rubio-Fernandez, 2019; Okanda et al., 2015; Panzeri & Foppolo, 2021). Thus we meticulously annotated each conversation to determine which of the four maxims in the cooperative principle have been violated.

In this work, we ask the following research questions:

(1) How do state-of-the-art LLMs perform in interpreting implicatures written in Chinese, in multiple choice style? (Section 4)

(2) What are the error patterns of LLMs on multiple choice questions? (Section 4)

(3) How do human participants rate the interpretation of conversational implicature generated by LLMs? (Section 5)

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1270–1285, Taiyuan, China, July 25 – 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

| Dataset | Task | Context | Manually | Language |
|---|---|---|---|---|
| CoQa (Reddy et al., 2019) | conversational QA | Paragraph | ✓ | English |
| Narrative-QA (Kočiský et al., 2018) | Reading comprehension | Documents | ✓ | English |
| DREAM (Sun et al., 2019) | QA | Free-form, multi-turn dialogue | ✓ | English |
| MuTual (Cui et al., 2020) | Next utterance prediction | Free-form, multi-turn dialogue | ✓ | English |
| GRICE (Zheng et al., 2021) | Gricean implicature recovery, QA | Fixed form, multi-turn dialogue | ✓ | English |
| PragMega (Floyd et al., 2023) | ToM QA | Paragraph | ✓ | English |
| DiPlomat (Li et al., 2023b) | Pragmatic identification and reasoning, QA | Free-form, multi-turn dialogue | ✗ | English |
| LUDWIG (Ruis et al., 2022) | Implicature QA | Single-turn dialogue | ✓ | English |
| Douban (Wu et al., 2017) | Next utterance prediction | Free-form, multi-turn dialogue | ✗ | Chinese |
| SwordsmanImp (This work) | Gricean Implicature QA | Free-form, multi-turn dialogue | ✓ | Chinese |

Table 1: Comparing our datasets and the existing datasets. "Manually" indicates whether the questions or answers are written partly or entirely by human.

We first review relevant research in Section 2 and then describe how we built our dataset SwordsmanImp in Section 3. In Section 4, we measure the models' accuracy on multiple-choice questions derived from our dataset. We test four models from GPT series (Radford et al., 2018), four open-source models. In Section 5, we use five models from Experiment 1 to generate pragmatic interpretations of 32 dialogues, and then we invite human participants to rate these interpretations in three dimensions: reasonability, logic and fluency.

We find that GPT-4 attains a human-level accuracy in multiple-choice questions. CausalLM (14B) also reaches a high accuracy as 78.5% on this task. Moreover, our results show that LLMs' performance does not vary with respect to different conversational maxims. Furthermore, we reveal that while in general fluent and self-consistent, occasional confused logic, redundant information and unexpected language tokens are the main reasons for their explanations to be underrated by human evaluators.

## 2 Related Work

Understanding non-literal meanings has long been considered a difficult task for language models. Previous studies have explored the capabilities of language models in recognizing metaphors (Wachowiak & Gromann, 2023; Neidlein et al., 2020), humor (Hessel et al., 2023; Jentzsch & Kersting, 2023; Chen et al., 2023) and social commonsense (Sap et al., 2019). Broad-scale datasets aimed at pragmatic reasoning collect data mainly through crowdsourcing or crawling from the internet (Reddy et al., 2019; Kočiský et al., 2018). Some others transform the existing tests for humans to train and evaluate language models (Cui et al., 2020; Sun et al., 2019). While the above-mentioned datasets have comprehensive coverage for pragmatic understanding, they are not ideal resources for evaluation of a specific type of pragmatic knowledge, that is, conversational implicatures derived by violation of the Gricean maxims. In our work, we craft our dataset manually to make sure that each implicature can be classified to one or more Gricean maxims. This enables us to perform a fine-grained analysis of LLMs' understanding of this particular pragmatic inference.

Previous attempts in evaluating LLMs' pragmatic understanding guided by the Gricean maxims or cooperative principles do not have multi-turn dialogues from real-world situations. Formalized under the theory of cooperative principles (Grice, 1975), Zheng et al. (2021) automatically generated their dataset GRICE through a set of well-defined grammar rules. This grammar-based method ensures a good control for the pragmatic cues to be aligned with the Gricean maxims. However, dialogues with fixed syntactic structures hardly represent the complexity of conversation in daily communication. In this study, we fill the gap by providing the multi-turn dialogue directly to the model.

Last but not least, there is a scarcity in Chinese evaluation resources for pragmatic inferences. Previous large-scale conversational datasets in Chinese use text sources from Sina Weibo (Shang et al., 2015), Douban conversational corpus (Wu et al., 2017) and E-commerce Dialogue corpus (Zhang et al., 2018). However, all of them are not specifically aimed to test LLMs' pragmatic understanding. To address this gap, we select the Chinese sitcom *My Own Swordsman* as the source to formulate pragmatic questions

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1270-1285, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

and construct a Chinese dataset for evaluating LLMs' pragmatic reasoning ability.

## 3 Dataset Construction

### 3.1 Data source

It is not easy to find naturally occurring, high-quality, multi-turn dialogues. Following previous literature that uses situational comedies (sitcom) as sources (Wang et al., 2017; Wu et al., 2021; Patro et al., 2021), we chose the Chinese sitcom *My Own Swordsman* (武林外传) as our data source, because of its abundance of conversational implicature, well-written dialogues and the uniquely Chinese background. Set in the Ming dynasty of China, this sitcom narrates the daily life of a group of people working in a Chinese motel. We believe it will be a unique resource for evaluating LLMs' pragmatic inference ability in both the Chinese language and the Chinese context.

### 3.2 Annotation procedure

**Implicature identification and classification** Three authors of this paper with at least two years of linguistics training went through the script of the sitcom to select conversations that contain conversational implicatures. Conversations with multiple turns were selected based on whether any conversational maxim is violated on purpose (Grice, 1975). For each conversation, we performed the cancellation test[1] to ensure that a conversational implicature rather than a semantic entailment is involved (Hirschberg, 1985). Multiple turns were included for each data entry to ensure that even if one has not seen the sitcom, one could still understand the conversation just from the snippet we selected.

Then, the same three authors classified all chosen dialogues according to the conversational maxims they violate. The criteria of these maxims are drawn from Grice (1975). To allow for a more fine-grained classification, we employ sub-maxims as criteria, assessing if the target sentence fulfills each requirement individually. An utterance is considered to have violated a maxim if it infringes on any sub-maxim. Moreover, a dialogue might belong to multiple classes according to the sub-maxims the utterance violates. An illustrative dataset entry featuring a dialogue, four interpretations and a class is presented in Figure 1.

**Writing four interpretations** Next, we construct the four interpretations of the sentence that carries implicature as the four choices: the pragmatic interpretation (the correct one), the literal interpretation, and two distractors with interpretations related to the context.[2] The pragmatic meaning are provided based on human commonsense understanding. We construct the literal meaning by rephrasing the target sentence with concrete and unambiguous expressions. We generate incorrect inferences as distractors, which are relevant to the the topic of the conversation. The character names referred to in the four interpretations are aligned with the mentions in the dialogue clip.

**Verification** We hired three PhD students in Linguistics to complete a multiple-choice question task, which is to choose the pragmatic understanding of the sentence from the shuffled four interpretations for all dialogues in our dataset. The students are required to complete the task independently without a time limit. After they finish, we invite them to discuss about their wrong answers and their reasoning process. This validation process guarantees that the provided pragmatic interpretation is closely aligned with the commonsense intuitive understanding and can be deduced from the limited context. Additional information that is necessary to pragmatic reasoning, such as relationships between characters, their personality, social background, and multi-modal information is supplemented in the bracket at the beginning of the dialogue. Besides, implicatures derived from vernacular, slang, and network-specific jokes have been filtered out from our dataset.

[1] A test to diagnose the conversational implicature by encoding semantically the negation of the target meaning. If the result seems consistent, then the target meaning is likely an implicature.

[2] The distractors can be understood as "neutral" statements in the Natural Language Inference task (Bowman et al., 2015).

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1270-1285, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

| | Total | Quality | Quantity | Relevance | Manner |
|---|---|---|---|---|---|
| # of questions | 200 | 76 | 33 | 71 | 62 |
| # of turns per dialogue | 6.80 | 7.84 | 5.91 | 6.23 | 6.35 |
| Avg. dialogue length | 158.22 | 184.53 | 143.67 | 147.20 | 152.79 |
| Avg. Utterance length | 23.27 | 23.53 | 24.31 | 23.64 | 24.04 |
| Avg. Answer length | 15.08 | 14.47 | 14.85 | 15.82 | 14.86 |

Table 2: Question numbers, average number of Chinese characters contained in each dialogue and utterance, and the number of turns per dialogue in our dataset.

### 3.3 Obtaining human score

To compare the results with human performance, we ask 10 native speakers to work on 32 questions randomly sampled from the dataset and they achieve an average accuracy of 93.1%. The questionnaire include the same number of each type of questions (i.e. Gricean maxims violated in the dialogue). The participants were all undergraduate students from a Chinese University, and they were compensated for their involvement in the experiment. Only 32 questions were used when estimating human score, because asking each annotator to solve 200 questions is implausible. The human accuracy should thus be regarded as an approximation of human performance.

### 3.4 Resulting corpus: SwordsmanImp

The final SwordsmanImp corpus contains 200 manually curated questions across four types categorized according to Gricean maxims. Each entry contains a multi-turn dialogue and four interpretations of the target sentence as choices (see Table 2).

## 4 Experiment 1: multiple-choice question for LLMs

In this experiment, the models will see the dialogue and the four interpretations we manually created. The task is to choose the correct interpretation (i.e., the pragmatic one) of the utterance that contains an implicature.

### 4.1 Models

We experiment with eight models in this experiment. The first four models are from the OpenAI GPT family, which are tested through the OpenAI API: text-davinci-002, text-davinci-003, GPT-3.5-turbo and GPT4.[3] We also examine four open-source models using the Transformers library (Wolf et al., 2020) from Huggingface[4]. We evaluate Chinese-Alpaca-2-13B (Cui et al., 2023), OpenBuddy-Llama2-13B[5] (based on Llama2 (Touvron et al., 2023)), CausalLM-13B (based on Llama2 and Qwen (Bai et al., 2023)), as well as BLOOMZ-7.1B from the BLOOM series (Muennighoff et al., 2023).

### 4.2 Evaluation Protocol

For both open-source and close-source models, we use zero-shot prompts to simulate the real-life situations where humans come across these non-literal expressions daily. However, we use two evaluation protocols depending on whether we have access to the probability/logits of all tokens in the vocabulary at a given output position. For close-source models, we ask it to generate the answer and manually go over the generated text to determine which interpretation is chosen. For open-source models, we follow established practices in LLM evaluation (Hu et al., 2023b; Hendrycks et al., 2020; Li et al., 2023a) where we compute the logits of the four tokens ("A", "B", "C", "D") after "答案："(answer:), and choose the one with the highest logit as model prediction (see an illustration in Appendix D.).

[3]The four OpenAI models are evaluated on November 15th, 2023.
[4]https://huggingface.co
[5]https://huggingface.co/OpenBuddy/openbuddy-llama2-13b-v8.1-fp16

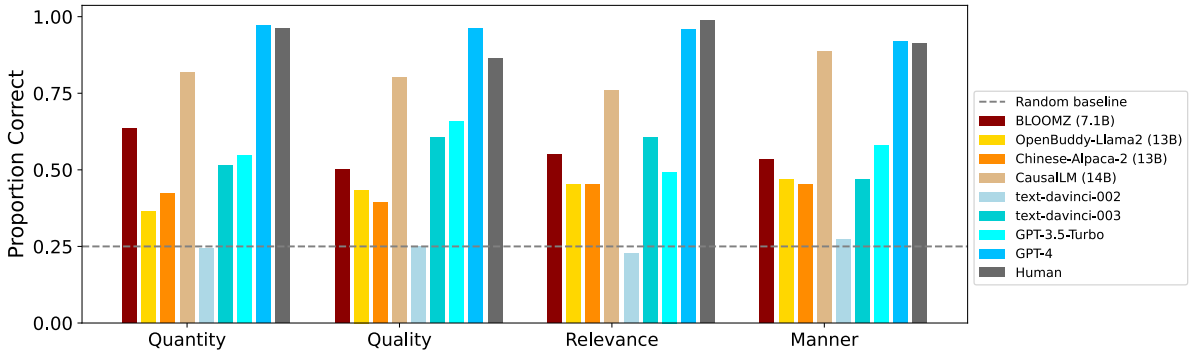



Figure 2: Performance of models across question types in the multiple choice scenario. Dashed line represents chance accuracy

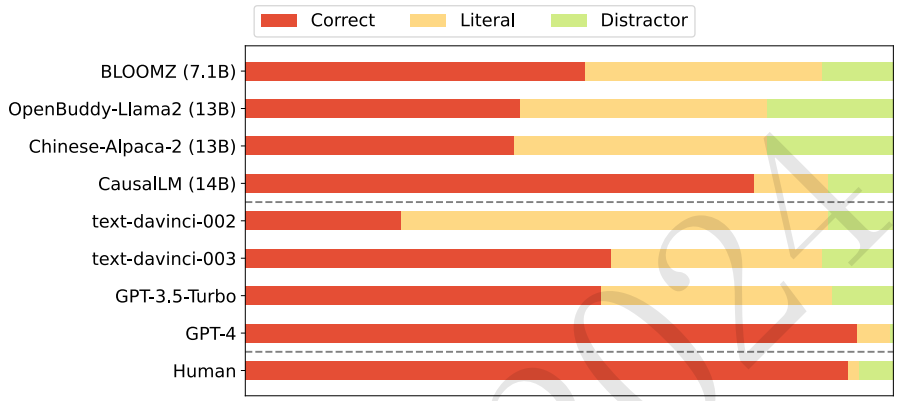



Figure 3: Frequency of each option in models' answers.

### 4.3 Results

The overall accuracy of all models and human participants is shown in Table 3. We observe that GPT-4 achieves the highest accuracy of 94%, showing a performance on par with human participants. CausalLM (14B) follows with 78.5% accuracy. Other models exhibit difficulty in identifying the pragmatic meaning of the line containing an implicature, with accuracy ranging from 20% to 60%. Text-davinci-002 even fails to perform above chance (25%). **This suggests that for most models we tested, there is still room for improvement in their understanding of implied meaning.** [6]

Table 2 shows that models' performance grouped by the Gricean maxims violated in the dialogue. Overall, certain models demonstrate proficiency in answering questions related to a set of maxims, while others excel in different ones. **We do not observe a uniform pattern indicating a particular strength or weakness in any of these maxims across the models.** This lack of uniformity is also observed in the results obtained from human participants who worked on the 32 sampled questions. Specifically, performance of text-davinci-002 is near to chance in all four conditions, demonstrating its incapacity in interpreting conversational implicature. The performance of OpenBuddy-Llama2 (13B) and Chinese-Alpaca-2 (13B) are comparable, both below 50% accuracy, with the two models demonstrate predominance in dealing with different types of questions. The accuracy of GPT-3.5-Turbo, text-davinci-003 and BLOOMZ (7.1B) are in the same range, and they also demonstrate different performance orders in different conditions. It's noteworthy that BLOOMZ (7.1B) outperforms the two 13B models in all conditions. CausalLM (14B) demonstrates an accuracy close to human-level and the best performance within all

[6]We also evaluate Baichuan2-13B-Chat and InternLM-Chat-20B (in half precision) with evaluation paradigm for close-source models. Their accuracy are separately 43% and 62%.

| Subject | Acc (%) |
|---|---|
| Random baseline | 25 |
| BLOOMZ (7.1B) | 52.5 |
| OpenBuddy-Llama2 (13B) | 42.5 |
| Chinese-Alpaca-2 (13B) | 41.5 |
| CausalLM (14B) | 78.5 |
| text-davinci-002 | 24.0 |
| text-davinci-003 | 56.5 |
| GPT-3.5-Turbo | 55.0 |
| GPT-4 | 94.0 |
| Human | 93.1 |

Table 3: Accuracy of language models and humans in experiment 1: multiple choice questions.

| | Reasonability | Logic | Fluency | Avg. response length |
|---|---|---|---|---|
| GPT-4 | $4.24 \pm 0.68$ | $4.65 \pm 0.39$ | $4.91 \pm 0.13$ | 114.44 |
| GPT-3.5-Turbo | $3.17 \pm 1.30$ | $4.09 \pm 0.77$ | $4.86 \pm 0.21$ | 125.41 |
| Chinese-Alpaca-2 (13B) | $2.34 \pm 1.10$ | $3.45 \pm 0.82$ | $4.72 \pm 0.39$ | 156.19 |
| CausalLM (14B) | $2.33 \pm 1.03$ | $3.48 \pm 0.67$ | $4.13 \pm 1.01$ | 147.41 |
| Openbuddy-Llama2 (13B) | $2.11 \pm 0.99$ | $3.55 \pm 0.71$ | $4.52 \pm 0.65$ | 153.56 |

Table 4: Models' mean scores in three dimensions with standard deviation and the average number of Chinese characters in their responses.

open-source models tested in this experiment. GPT-4 attains the highest accuracy of 94.0%, surpassing 90% across all categories of questions. GPT-4's performance, although slightly higher than the human score, is not statistically different from the human accuracy ($p = 0.802 > 0.05$ by a two-tailed t-test).

The distribution of the interpretations chosen by the models is shown in Figure 3, where red means that the model has chosen the correct answer, i.e., pragmatic interpretation, while yellow corresponds to the literal meaning, and green the two distractors. The performance of two 13B models shows a higher frequency of choosing distractors. This could possibly indicate a that these two models are easily sidetracked by irrelevant information in the context. We also observe that as the GPT models evolve, they have a higher chance of distinguishing literal meaning from implied meaning, culminating in the considerably low ratio of literal understanding in GPT-4.

## 5 Experiment 2: evaluating the quality of explanations generated by LLMs

In the previous experiment, we ask LLMs to choose one answer from four choices. In this experiment, we design open-ended questions where the models are asked to generate explanations of the implicature, which will then be evaluated manually by native speakers of Chinese, based on the reasonability/reasonableness, logic and fluency of the generated explanations.

### 5.1 Experimental setup

We first performed a pilot trial comprising five questions. Among the eight models, BLOOMZ (7.1B), text-davinci-002 and text-davinci-003 produce short and fragmented responses, despite our prompts explicitly asking for detailed explanation. We therefore select GPT-3.5-Turbo, GPT-4, Chinese-Alpaca-2 (13B), OpenBuddy-Llama2 (13B) and CausalLM (14B) for this experiment. We increase `max_new_tokens` for open-sourced models, and keep the same hyperparameter for GPT models (See Appendix B). We use the prompt in Figure 4 to instruct the models to formulate an explanation for the target sentence. This prompt is selected after a series of trial experiments to ensure the quality of models' responses.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1270-1285, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

| Question | Model Responses |
|---|---|
| 你现在是一个中文母语者。对于以下对话，请识别特定人物的话语中的言外之意，并解释。<br>**李大嘴：** 是啊，要么怎么说是龙门镖局的少东家呢，一般人哪有这气派是不是。啥叫家传，啥叫传统（刚说道这里佟石头“哇”的一声吐了出来，众人挥手趋散气味）<br>**佟湘玉：** 这可不是我家的传统。<br>**佟石头：** 再来！<br>**佟湘玉：** 佟石头你有完没完了？<br>**佟石头：** 给洒家倒酒！<br>**佟湘玉：** 不用倒了，捧着坛子一口气喝完那多爷们儿啊！<br>**佟石头：** 姐，那一坛子二斤半呐！<br>**佟湘玉：** 你不是一口气能喝五六十斤嘛？<br>请根据以上情景判断**佟湘玉**说的“你不是一口气能喝五六十斤嘛？”有什么言外之意，并解释。<br><br>You are now a native Chinese speaker. Regarding the following conversation, please identify the implicit meanings in the speech of specific individuals and explain them.<br>Dazui : Yeah, otherwise, how could he be considered the young master of the Longmen Escort Agency[1]? Ordinary people don't have this aura, right? What's meant by family heritage, what's tradition...(Just as he was saying this, Shitou, with a "wow," spat out, and everyone waved their hands, dispersing the smell.)<br>Xiangyu : That's not our family's tradition.<br>Shitou : Next!<br>Xiangyu : Shitou, are you ever going to finish?<br>Shitou : Pour me a drink!<br>Xiangyu : No need to pour. If you hold the jug, you'll finish it all, big guy!<br>Shitou : Sis, that jug is fourty four ounces!<br>Xiangyu : Aren't you capable of drinking a thousand ounces in one gulp?<br>Please judge the implicit meaning of Xiangyu's statement "Aren't you capable of drinking a thousand ounces in one gulp?" based on the above context, and explain it. | **GPT-4:** 佟湘玉的话”你不是一口气能喝五六十斤嘛？”的言外之意是在讽刺和挑衅佟石头。她的话语表面上是在赞扬佟石头的饮酒能力，实际上是在质疑他的酒量，挑衅他是否真的能一口气喝完一整坛的酒。同时，这句话也表达了她对佟石头酒后行为的不满和无奈。<br>(114 characters)<br>Xiangyu's statement "Aren't you capable of drinking a thousand ounces in one gulp?" implies sarcasm and provocation towards Shitou. On the surface, her words seem to praise Shitou's drinking ability, but in reality, she questions his capacity for alcohol, challenging whether he can really drink an entire jug in one go. Simultaneously, this sentence also conveys her dissatisfaction and helplessness toward Shitou's behavior when he's drunk.<br><br>**CausalLM (14B):** 佟湘玉的言外之意是质疑佟石头的酒量，暗示他可能酒品NST。她可能认为，佟石头的言语和行为过于狂妄自大，让他在众人面前出丑。通过这句话，佟湘玉试图让佟石头意识到自己的错误，并提醒他不要forgot his place.<br>(109 characters)<br>Xiangyu's implicit meaning questions Shitou's alcohol capacity, suggesting he might not handle alcohol well NST. She might perceive Shitou's words and actions as overly arrogant, causing embarrassment in front of others. Through this statement, Xiangyu attempts to make Shitou aware of his mistake and reminds him not to forget his place.<br><br>**OpenBuddy-Llama2 (13B):** 从这段对话中，我们可以看出佟湘玉说的话语中存在一些言外之意。她说“你不是一辈子能喝五六八番呢？”其实这句话暗示着佟芒头喝酒的行为可能已经到了一定的程度，而她自己也已经接受了这个现实。这种暗示的意思是，佟芒头喝酒的习惯已经形成，他已经不能像以前那样随意地停止喝酒了。因此，佟湘玉的话可以被解释为一种对佟芒头喝酒行为的默许和接受。<br>(163 characters)<br>From this conversation, we can see that Xiangyu's words contain some implicit meanings. When she says, "Aren't you capable of drinking for a lifetime?" in fact, this sentence implies that Shitou's drinking behavior may have reached a certain level, and she herself has accepted this reality. The implication here is that Shitou's drinking habit has formed, and he can no longer stop drinking as freely as before. Therefore, Xiangyu's words can be interpreted as a kind of tacit approval and acceptance of Shitou's drinking behavior. |

Figure 4: Models' interpretations to a character line (highlighted with red color) that carries extra meanings in a dialogue with multiple turns. Explanations aberrant from the correct interpretation and non-Chinese characters are highlighted with yellow bottom color. Escort Agency refers to historical security firms that were typically hired by trade merchants for protection during transportation of goods. They were known to use martial arts for defense, similar to bodyguards or security personnel.

## 5.2 Human evaluation

We invite 15 human participants, all of whom are linguistics majors at a top Chinese university, to evaluate the answers generated by models in three dimensions: reasonability or reasonableness, logic and fluency. Reasonability is gauged by the degree to which the response is aligned with the interpretation provided by the human annotator. Logic is assessed based on the presence of a clear and consistent chain of reasoning in the response. Fluency, on the other hand, evaluates whether the generated text is devoid of incoherent elements such as foreign language insertions or unintelligible code fragments.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1270-1285, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

### 5.3 Results

Table 4 shows the mean ratings of the responses of the five models' in three dimensions. Responses from GPT-4 scores the highest in all three dimensions with the smallest variance. Responses from GPT-3.5-Turbo are rated high but with larger standard deviations, suggesting unstable performance. Answers from the other three models are comparable to each other ($F = .964$, $p = .549$). Notably, the scores of CausalLM (14B) are lower than those of GPT-3.5-Turbo, which is inconsistent to results of Experiment 1. As providing coherent explanations of the implicatures is much harder than picking one answer from four options, this suggests that a model good at the latter may not necessarily be equally good at the former.

Examining the generated answers in detail explains the distribution of human rating scores. Text generated by GPT-4 and GPT-3.5-Turbo are both identified with a minimal number of "code switching", the phenomenon of adulterating non-Chinese tokens into their responses, with the interpretation from GPT-4 being more precise and coherent than that from GPT-3.5-Turbo. While Chinese-Alpaca-2 (13B), OpenBuddy-Llama2 (13B) and CausalLM (14B) have comparable performance, they all underperform the two GPT models. They reflect different patterns of generated text. Both Chinese-alpaca-2 (13B) and Openbuddy-Llama2 (13B) feature highly pure Chinese responses with their scores in text quality above 4.5, but the latter model scores distinctively low in reasonability. CausalLM-14B, on the contrary, tends to code-switch or generate tokens in English frequently.

Figure 4 presents a typical example to illustrate the different styles of generation from the models. The target sentence in this dialogue doesn't reflect a normal judgment based on his alcoholic words and behaviors but rather proposes a totally unreal scenario. The overt absurdity in the words represents an obvious signal of the violation of the maxim of Quality. An implicature from Xiangyu that Shitou can not drink anymore thus arises and is precisely conveyed to the listeners. Her words also express her irony and dissatisfaction towards Shitou. GPT-4 gives a concise interpretation that aligns closest to the reference interpretation among the selected models. However, it has mistaken the ironic tone as questioning Shitou's capacity for liquor, as highlighted in the example. CausalLM (14B) produces a correct interpretation by and large, but the quality of its answer is negatively impacted by its poor fluency, in that some English words and meaningless character sequences appear in its answer. It is interesting that the "forgot his place" carries the correct meaning, which leads us to consider this as code-switch, rather than nonsensical generation. This code-switching phenomenon appears to be a feature specific to the responses from CausalLM (14B). The response from Openbuddy-Llama2 (13B) exemplifies a response that is both verbose and irrelevant.

## 6 Discussion and Future Work

### 6.1 LLM's understanding of conversational implicature in Chinese?

Our results from Experiment 1 show that the performance of GPT4 on our proposed benchmark is on par with humans, while other models are at least 15 points behind (including GPT-3.5-turbo). This suggests that while in principle pragmatic implicatures can be acquired by arguable the best LLMs at the moment, it is a non-trivial task for other LLMs.

Results from Experiment 1 also reveal no significant by-maxim variance in human accuracy, as well as model accuracy (see Figure 2). This is different from the results in previous work on human processing of implicatures (Engelhardt et al., 2006; Rubio-Fernandez, 2019; Okanda et al., 2015; Panzeri & Foppolo, 2021), which demonstrate that humans sanction infringements of the maxims in different ways, being less sensitive to the violation of the maxim of quantity than to others, leading to more processing difficulty for this maxim. This difference is possibly because many dialogues in our dataset violate several maxims, rather than a single one. Thus we do not observe a difference in human processing of implicature derived from violation of different maxims.

Similar in evaluation paradigm, Hu et al. (2023a) classifies their pragmatic-related questions according to Theory of Mind (ToM), in which the violation of Gricean maxims is a single category without sub-categories. Text-davinci-002 is evaluated in both their and our studies, resulting in different performances. It surpasses 60% accuracy on the English dataset in Hu et al. (2023a), but does not reach 25% accuracy on our dataset. We attribute this to the difference in our selection of text sources. While Hu

CCL 2024

et al. (2023a) uses hand-crafted scenarios targeting humor, sarcasm, and other pragmatic phenomena separately, often with fewer turn-taking, our text source comes from the script of the sitcom, with many more turn-taking, each of which contains multiple phenomena. The degradation in text-davinci-002's performance can serve as a calibration of the difficulty of questions in the two datasets. With text-davinci-002 being the only model with accuracy lower than 40% in the current study, the results demonstrate a perceptible development in the pragmatic understanding of the newly developed LLMs.

### 6.2 Future directions

Results from Experiment 2 indicates the possibility that a model with high accuracy in multiple-choice question could fail in a free-text generation task of interpreting the pragmatic meaning on its own (CausalLM-14B). We thus argue that multiple-choice questions alone is not enough for a comprehensive evaluation of LLMs' linguistic ability. Manual inspection of free-form generation is a must for a more robust analysis of model performance. We also see the potential of a more sophisticated design to better quantify their free-form explanations of conversational implicature.

Future work can also create a large-scale dataset composed of conversational implicature embedded in *naturally occurring* dialogues. This will require detailed annotation of spoken corpora, which we believe could benefit from the procedure and implicature-selection criteria used in creating our benchmark.

## 7 Conclusion

In this paper, we present SwordsmanImp, the first fine-grained Chinese dataset to evaluate LLMs' understanding of conversational implicature. In two experiments, we evaluate the state-of-the-art language models' pragmatic skills with two tasks. In Experiment 1, we reveal that GPT-4 attains a human-level accuracy in answering multiple-choice questions, with other models lagging behind. We also find that no significant difference exists for both LLMs and human's accuracy with respect to conversational implicatures that violate different maxims. Results from Experiment 2 reflect the different patterns of generated text across LLMs and after human annotation of the quality of generated text from three dimensions, we reveal that while most models produce fluent text, they struggle to generate coherent and sensible explanations for the implicature, even if the model has achieved high accuracy in multiple-choice question.

## Acknowledgment

We thank Xinjia Qi, Qiyu Sun and Yaqian Zhang for verifying the implicatures and improving the dataset. We thank all participants for their support in this study. We also thank the anonymous reviewers for their valuable comments. This project is funded by Shanghai Pujiang Program (22PJC063) awarded to Hai Hu.

## Limitations

Our dataset is sourced exclusively from the Chinese sitcom *My Own Swordsman*. Although we performed a rigorous proofreading on our data, selecting dialogues whose interpretation only depends on the local information and providing background knowledge when necessary, there may still be features specific to this sitcom such as the personality of the characters that play a role in determining the implicature, which may influence the generalizability of our conclusion.

## References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 632–642, 2015.

Penelope Brown and Stephen C Levinson. *Politeness: Some universals in language usage*, volume 4. Cambridge university press, 1987.

Yuyan Chen, Zhixu Li, Jiaqing Liang, Yanghua Xiao, Bang Liu, and Yunwen Chen. Can pre-trained language models understand chinese humor? In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pp. 465–480, 2023.

Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. MuTual: A dataset for multi-turn dialogue reasoning. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1406–1416, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.130. URL https://aclanthology.org/2020.acl-main.130.

Yiming Cui, Ziqing Yang, and Xin Yao. Efficient and effective text encoding for chinese llama and alpaca, 2023.

Paul E Engelhardt, Karl GD Bailey, and Fernanda Ferreira. Do speakers and listeners observe the gricean maxim of quantity? *Journal of memory and language*, 54(4):554–573, 2006.

Sammy Floyd, Edward Gibson, and Evelina Fedorenko. Pragmega. Retrieved from https://osf.io/dpge6, September 2023.

Herbert P Grice. Logic and conversation. In *Speech acts*, pp. 41–58. Brill, 1975.

Herbert Paul Grice. Retrospective epilogue. *Studies in the Way of Words*, 339:386, 1989.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2020.

Jack Hessel, Ana Marasovic, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. Do androids laugh at electric sheep? humor "understanding" benchmarks from the new yorker caption contest. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 688–714, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.41. URL https://aclanthology.org/2023.acl-long.41.

J.L.B. Hirschberg. *A Theory of Scalar Implicature*. University of Pennsylvania, 1985. URL https://books.google.com/books?id=bvEYAQAAIAAJ.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. A fine-grained comparison of pragmatic language understanding in humans and language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4194–4213, Toronto, Canada, July 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.230. URL https://aclanthology.org/2023.acl-long.230.

Jennifer Hu, Roger Levy, Judith Degen, and Sebastian Schuster. Expectations over unspoken alternatives predict pragmatic inferences. *arXiv preprint arXiv:2304.04758*, 2023b.

Sophie Jentzsch and Kristian Kersting. Chatgpt is fun, but it is not funny! humor is still challenging large language models. *arXiv preprint arXiv:2306.04563*, 2023.

Zae Myung Kim, David E Taylor, and Dongyeop Kang. " is the pope catholic?" applying chain-of-thought reasoning to understanding conversational implicatures. *arXiv preprint arXiv:2305.13826*, 2023.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328, 2018. doi: 10.1162/tacl_a_00023. URL https://aclanthology.org/Q18-1023.

CCL 2024

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. CMMLU: Measuring massive multitask language understanding in chinese. *arXiv preprint arXiv:2306.09212*, 2023a.

Hengli Li, Song-Chun Zhu, and Zilong Zheng. Diplomat: A dialogue dataset for situated pragmatic reasoning, 2023b.

Benjamin Lipkin, Lionel Wong, Gabriel Grand, and Joshua B Tenenbaum. Evaluating statistical language models as pragmatic reasoners. *arXiv preprint arXiv:2305.01020*, 2023.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. Crosslingual generalization through multitask finetuning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15991–16111, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.891. URL https://aclanthology.org/2023.acl-long.891.

Arthur Neidlein, Philip Wiesenbach, and Katja Markert. An analysis of language models for metaphor recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 3722–3736, 2020.

Mako Okanda, Kosuke Asada, Yusuke Moriguchi, and Shoji Itakura. Understanding violations of gricean maxims in preschoolers and adults. *Frontiers in psychology*, 6:901, 2015.

Lalchand Pandia, Yan Cong, and Allyson Ettinger. Pragmatic competence of pre-trained language models through the lens of discourse connectives. *arXiv preprint arXiv:2109.12951*, 2021.

Francesca Panzeri and Francesca Foppolo. Children' s and adults' sensitivity to gricean maxims and to the maximize presupposition principle. *Frontiers in Psychology*, 12:624628, 2021.

Badri N Patro, Mayank Lunayach, Deepankar Srivastava, Hunar Singh, Vinay P Namboodiri, et al. Multimodal humor dataset: Predicting laughter tracks for sitcoms. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 576–585, 2021.

Zhuang Qiu, Xufeng Duan, and Zhenguang Cai. Does chatgpt resemble humans in processing implicatures? In *Proceedings of the 4th Natural Logic Meets Machine Learning Workshop*, pp. 25–34, 2023.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

Siva Reddy, Danqi Chen, and Christopher D Manning. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266, 2019.

Paula Rubio-Fernandez. Overinformative speakers are cooperative: Revisiting the gricean maxim of quantity. *Cognitive science*, 43(11):e12797, 2019.

Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. Large language models are not zero-shot communicators. *arXiv preprint arXiv:2210.14986*, 2022.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social IQa: Commonsense reasoning about social interactions. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4463–4473, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1454. URL https://aclanthology.org/D19-1454.

John R Searle, Ferenc Kiefer, Manfred Bierwisch, et al. *Speech act theory and pragmatics*, volume 10. Springer, 1980.

CCL 2024

Lifeng Shang, Zhengdong Lu, and Hang Li. Neural responding machine for short-text conversation. In Chengqing Zong and Michael Strube (eds.), *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1577–1586, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1152. URL https://aclanthology.org/P15-1152.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. DREAM: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231, 2019. doi: 10.1162/tacl_a_00264. URL https://aclanthology.org/Q19-1014.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Lennart Wachowiak and Dagmar Gromann. Does gpt-3 grasp metaphors? identifying metaphor mappings with generative language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1018–1032, 2023.

Xiaolong Wang, Rohit Girdhar, and Abhinav Gupta. Binge watching: Scaling affordance learning from sitcoms. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2596–2605, 2017.

Deirdre Wilson and Dan Sperber. Relevance theory. *The handbook of pragmatics*, pp. 606–632, 2006.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Jiaming Wu, Hongfei Lin, Liang Yang, and Bo Xu. Mumor: A multimodal dataset for humor detection in conversations. In *Natural Language Processing and Chinese Computing: 10th CCF International Conference, NLPCC 2021, Qingdao, China, October 13–17, 2021, Proceedings, Part I 10*, pp. 619–627. Springer, 2021.

Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 496–505, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1046. URL https://aclanthology.org/P17-1046.

Zhuosheng Zhang, Jiangtong Li, Pengfei Zhu, Hai Zhao, and Gongshen Liu. Modeling multi-turn conversation with deep utterance aggregation. In Emily M. Bender, Leon Derczynski, and Pierre Isabelle (eds.), *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 3740–3752, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL https://aclanthology.org/C18-1317.

Zilong Zheng, Shuwen Qiu, Lifeng Fan, Yixin Zhu, and Song-Chun Zhu. Grice: A grammar-based dataset for recovering implicature and conversational reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 2074–2085, 2021.

## A Evaluation paradigms and differences

Models' performance might vary from how their answers are estimated and collected. Next token prediction and free generation are two paradigms separately used to estimate the performance of open-source

| Model | Accuracy (%) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Total | | Quantity | | Quality | | Relevance | | Manner | |
| | Next | Gen | Next | Gen | Next | Gen | Next | Gen | Next | Gen |
| BlOOMZ | 52.50 | 35.50 | 63.64 | 24.24 | 50.00 | 31.58 | 54.93 | 30.99 | 53.23 | 33.87 |
| OpenBuddy-Llama2 | 42.50 | 21.50 | 36.36 | 21.21 | 43.42 | 25.00 | 45.07 | 19.72 | 46.77 | 17.74 |
| Chinese-Alpaca-2 | 41.50 | 42.00 | 42.42 | 36.36 | 39.47 | 44.74 | 45.07 | 43.66 | 45.16 | 38.71 |
| CausalLM | 78.50 | 31.50 | 81.82 | 42.42 | 80.26 | 31.58 | 76.06 | 30.99 | 88.71 | 40.32 |

Table 5: Comparison between the accuracy of open-source models on multiple-choice questions when evaluated with next token prediction and free generation paradigms.

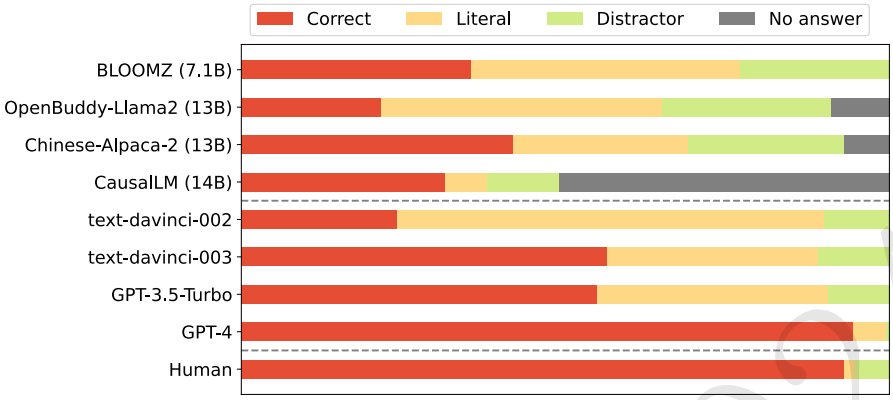



Figure 5: Answer distribution of models when answers are collected with free generation paradigm.

and close-source models in Experiment 1. Table 5 shows the comparison of the open-source models' performance on the multiple-choice questions when their answers are estimated through the two paradigms. The result shows a decrease in accuracy in BLOOMZ (7.1B), CausalLM (13B), and OpenBuddy-Llama2 (13B) and a slight increase in Chinese-Alpaca-2 (13B) when the paradigm switches from next token prediction to free generation. This is aligned with the findings of Li et al. (2023a). Among the four models CausalLM (13B) has a dramatic decrease in its accuracy, from 78.5% to 31.5%, which corresponds to its bad performance in Experiment 2. We find that it fails to give a definite answer in its responses for over half of the questions, as shown in Figure 5.

## B Hyperparameter setting

The hyperparameters used for gathering responses from open-source models in Experiment 1 and Experiment 2 are shown in Table 6 and Table 7.

CCL 2024

| Parameter | Value |
|---|---|
| max_new_tokens | 50 |
| temperature | 0.9 |
| top_k | 3 |
| top_p | 0.1 |
| repetition_penalty | 1.0 |
| num_return_sequence | 1 |
| do_sample | True |

Table 6: Parameter setting for open-source models in Experiment 1

| Parameter | Value |
|---|---|
| max_new_tokens | 300 |
| temperature | 0.9 |
| top_k | 0 |
| top_p | 0.9 |
| repetition_penalty | 1.0 |
| num_return_sequence | 1 |
| do_sample | True |

Table 7: Parameter setting for open-source models in Experiment 2

## C Average answer length for different question types

We present the average answer length for the four maxims in Table 8.

## D Prompt for experiment 1

你现在是一个中文母语者。对于以下对话，请识别特定人物的话语中的的言外之意，在给出的四个选项中选择一个你认为的正确答案。[En: You are now a native Chinese speaker. For the following dialogue, please identify the implied meaning in the specific character's speech, and choose one of the four given options that you think is the correct answer. ]

< Dialogue >

< Four interpretations as choices >

(Close-source models:) 请在 'Response:' 后写出你选择的答案。[En: Please write your answer after 'Response:']

(Open-source models:) 答案：[En: Answer:]

CCL 2024

| | Pragmatic | Literal | Distractor#1 | Distractor#2 |
|---|---|---|---|---|
| Total | 19.20 | 16.38 | 12.54 | 12.21 |
| Quality | 18.32 | 15.88 | 11.96 | 11.72 |
| Quantity | 18.36 | 16.73 | 12.30 | 12.00 |
| Relevance | 20.77 | 16.62 | 13.11 | 12.77 |
| Manner | 18.10 | 17.18 | 12.13 | 12.06 |

Table 8: Average number of Chinese characters in each answer for all questions and for each type of questions.

CCL 2024

# EmoFake: An Initial Dataset for Emotion Fake Audio Detection

**Yan Zhao**[1,2], **Jiangyan Yi**[2], **Jianhua Tao**[3], **Chenglong Wang**[2], **Yongfeng Dong**[1]✉
[1]School of Artificial Intelligence, Hebei University of Technology, China
[2]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, China
[3]Department of Automation, Tsinghua University, China
202122802007@stu.hebut.edu.cn;jiangyan.yi,chenglong.wang@nlpr.ia.ac.cn
jhtao@tsinghua.edu.cn;dongyf@hebut.edu.cn

## Abstract

To enhance the effectiveness of fake audio detection techniques, researchers have developed multiple datasets such as those for the ASVspoof and ADD challenges. These datasets typically focus on capturing non-emotional characteristics in speech, such as the identity of the speaker and the authenticity of the content. However, they often overlook changes in the emotional state of the audio, which is another crucial dimension affecting the authenticity of speech. Therefore, this study reports our progress in developing such an emotion fake audio detection dataset involving changing emotion state of the origin audio named EmoFake. The audio samples in EmoFake are generated using open-source emotional voice conversion models, intended to simulate potential emotional tampering scenarios in real-world settings. We conducted a series of benchmark experiments on this dataset, and the results show that even advanced fake audio detection models trained on the ASVspoof 2019 LA dataset and the ADD 2022 track 3.2 dataset face challenges with EmoFake. The EmoFake is publicly available[1] now.

## 1 Introduction

In the past few years, Voice conversion (VC) technology has been able to generate very natural converted audio. But it is still not equipped with human-like emotions adequately. Emotion, as a vital component of human communication, plays a crucial role in manifesting itself on the semantic and pragmatic levels of spoken language. Therefore, the research of emotion voice conversion (EVC) has emerged. Generally speaking, VC refers to convert the speaker's identity while retaining linguistic information. AS a special type of VC, EVC converts the emotion state of the speech from the source emotion to the target emotion as shown in Figure 1.

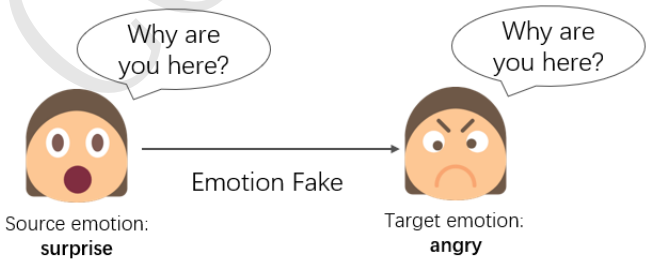



Figure 1: The emoiton fake audio changed the source emotion like surprise into the target emotion like angry without changing identity of speaker, content or other information.

Currently, the framework commonly used on EVC task are: VAW-GAN (Zhou et al., 2020b; Zhou et al., 2021b), Sequence-to-Sequence (Seq2Seq) (Zhou et al., 2021a), CycleGAN (Zhou et al., 2020a; Fu et al., 2022; Gao et al., 2019) and StarGAN (Rizos et al., 2020). In personalized voice generation,

[1]https://zenodo.org/records/10443769

©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1286-1297, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers

such as singing voice conversion, and intelligent conversational systems like artificial intelligence voice assistants, EVC plays an indispensable role. However, if people misuse this technology to attack the security system or interfere with the forensic process in a case, it can pose the immeasurable impact on people's lives and social stability. In a legal context, imagine a crucial piece of evidence is a recorded confession. Using EVC technology, someone could manipulate the recording by altering the suspect's emotional tone. This could lead to a misinterpretation of the confession, casting doubt on the reliability of the evidence and influencing the case outcome. Therefore, enhancing the ability to discriminate emotion fake audio is urgently needed.

There has been a lot of very interesting work on the fake audio detection. Comprehensive datasets have been developed to improve the performance of the fake audio detection model. Based on SAS (Wu et al., 2015a), the first Automatic Speaker Verification Spoofing and Countermeasures Challenge (ASVspoof) (Wu et al., 2015b) have been organized in 2015. The dataset of ASVspoof 2017 (Kinnunen et al., 2017) promoted the development of replay attack countermeasures. The dataset of ASVspoof 2019 (Todisco et al., 2019) consists of text to speech (TTS), VC, and replay attacks. Compared to previous datasets, the dataset of ASVspoof 2021 (Yamagishi et al., 2021) focus upon deepfake speech in addition. Datasets of the first Audio Deep synthesis Detection (ADD 2022) (Yi et al., 2022) and the second Audio Deep fake Detection (ADD 2023) (Yi et al., 2023a) add partially fake audio (Yi et al., 2021) to further accelerate and foster research on detecting deepfake audio. Yi et al. (Yi et al., 2024) developed a dataset which consists of the scene fake audio. Yi et al. (Yi et al., 2023b) highlight the key differences across various types of deepfake audio, then outline and analyse competitions, datasets, features, classifications. Zang et al. (Zang et al., 2024) proposed the singing voice deepfake detection task and presented a database named SingFake. Müller et al. proposed the MLAAD (Müller et al., 2024) dataset, created using 54 TTS models in 23 different languages. These datasets have been developed to further the development of research on fake audio detection. Fake audio in previous dataset is mainly generated by changing timbre, linguistic content, channel noise of the original utterance or acoustic scene. However, none of previous datasets consider a scenario where the emotional state of speech is altered.

Therefore, to enhance the discrimination ability of the fake audio detection model against the emotion fake audio, this paper reports our progress in developing an emotion fake audio detection dataset involving changing emotional state in the speech. We developed a dataset named EmoFake consisting the emotion fake audio and real emotional speech, which is publicly available[1]. We also provided some benchmark results for detecting the emotion fake audio. The results show that our designed dataset poses a challenge to the fake audio detection model trained with the LA dataset of ASVspoof 2019. Existing mainstream foundational fake audio detection systems experience varying degrees of performance degradation when faced with the emotion fake audio.

The rest of this paper is organized as follows. Section 2 describes the design policy of the EmoFake. Section 3 presents the experiments and baselines. Section 4 concludes this work.

## 2 Dataset design

EmoFake is based on the dataset named Emotional Speech Database (ESD) (Zhou et al., 2022) which is designed for multi-speaker and cross-lingual EVC studies[1]. ESD includs five male English speakers, five female English speakers, five male Chinese speakers, and five female Chinese speakers. Each speaker spoke 350 parallel utterances using five emotions: Neutral (Neu), Happy (Hap), Angry (Ang), Sad and Surprise (Sur), and the average duration of each utterance was 2.9 seconds. Each utterance is recorded in a standard room environment at a sampling rate of 16 kHz and on a signal-to-noise ratio of 20 dB or more.

### 2.1 Fake audio generation

As of now, there have been numerous excellent works of EVC, broadly categorized into two types: method based on Generative Adversarial Networks (GANs) and methods based on Seq2Seq models. We used all the representative open-source EVC models to generate the emotion fake audio in EmoFake:

[1]https://github.com/HLTSingapore/Emotional-Speech-Data

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1286-1297, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Table 1: The description of EVC models used in generating the emotion fake audio in EmoFake.

| Framework | ID | Model name | Description |
|---|---|---|---|
| VAW-GAN | S1 | VAW-GAN-CWT (Zhou et al., 2020b) | A speaker-independent framework for EVC which uses a encoder structure based on the VAW-GAN. It performs the continuous wavelet transform (CWT) of fundamental frequency (F0) to describe the prosody from micro-prosody level to utterance level. |
| | S2 | DeepEST (Zhou et al., 2021b) | It uses a pre-trained speech emotion recognition (Chen et al., 2018) system to extract deep emotion features. This EVC model achieves one-to-many emotion conversion and can handle unseen emotions. |
| Seq2Seq | S3 | Seq2Seq-EVC (Zhou et al., 2021a) | A Sequence-to-Sequence framework for EVC with a two-stage training strategy. In the first stage, the style encoder is pretrained using a multi-speaker TTS corpus. In the second stage, the emotion encoder is trained using emotional speech data. |
| CycleGAN | S4 | CycleGAN-EVC (Zhou et al., 2020a) | A parallel-data-free EVC framework that performs both spectrum and prosody conversion using CycleGAN and CWT. |
| | S5 | CycleTransGAN (Fu et al., 2022) | Leveraging the transformer (Vaswani et al., 2017), the model expands its receptive field, enabling the generated speech to exhibit greater temporal feature consistency. This helps mitigate issues related to mispronunciations and skipped phonemes, addressing instability problems to some extent. |
| | S6 | EmoCycleGAN (Gao et al., 2019) | A nonparallel EVC approach based on style transfer autoencoders and CycleGAN based framework. |
| StarGAN | S7 | StarGAN-EVC (Rizos et al., 2020) | The harmonic frequency contains emotion indices, StarGAN is used to learn how to transform the spectral envelope with aperiodicity parameters, the 36 cepstral coefficients and the F0 contour. |

VAW-GAN-CWT[2], DeepEST[3], Seq2Seq-EVC[4], CycleGAN-EVC[5], CycleTransGAN[6], EmoCycleGAN[7] and StarGAN-EVC[8]. The detailed description of the EVC models we used are shown in Table 1.

In order to ensure EVC models can achieve the highest performance, when training each EVC model, the English EVC model is trained using the utterances of ESD Chinese and English speakers, and the Chinese EVC model is trained using the utterances of Chinese speakers. When generating emotion fake audio, the EVC model of different languages generated emotion fake audio in the corresponding language. Furthermore, we believe that enabling each EVC model to perform emotion voice conversion on both male and female speakers' speech can make the fake audio in EmoFake more comprehensive. Therefore, all twenty speakers are divided into ten pairs, with each pair consisting of one male speaker and one female speaker of the same language. The information of the speaker pair is shown in Table 2. To maximize the diversity of emotion conversion pairs, we selected different source emotions for various

[2]https://github.com/KunZhou9646/Speaker-independent-emotional-voice-conversion-based-on-conditional-VAW-GAN-and-CWT

[3]https://github.com/KunZhou9646/controllable_evc_code

[4]https://github.com/KunZhou9646/seq2seq-EVC

[5]https://github.com/KunZhou9646/emotional-voice-conversion-with-CycleGAN-and-CWT-for-Spectrum-and-F0

[6]https://github.com/CZ26/CycleTransGAN-EVC

[7]https://github.com/bottlecapper/EmoCycleGAN

[8]https://github.com/glam-imperial/EmotionalConversionStarGAN

Table 2: The description of speaker pairs.

| Language | Group ID | Male speaker ID | Female speaker ID |
|---|---|---|---|
| Chinese | C1 | 0006 | 0001 |
| | C2 | 0004 | 0003 |
| | C3 | 0005 | 0002 |
| | C4 | 0010 | 0009 |
| | C5 | 0008 | 0007 |
| English | E1 | 0011 | 0016 |
| | E2 | 0013 | 0017 |
| | E3 | 0012 | 0015 |
| | E4 | 0020 | 0019 |
| | E5 | 0014 | 0018 |

speaker pairs. For each pair of speakers, we converted their source emotional speech data in ESD into utterances containing each of the remaining four emotions.

### 2.2 Dataset composition

According to the language type in ESD, EmoFake is divided into two subsets: Chinese subset and English subset. Each subset is divided into three parts: train, dev and test. The train part includes three pairs of speakers, while each of dev each and test part include one pair of speakers. For each speaker pair, the speech of all five emotions of two speakers in the ESD dataset is regarded as the real audio of EmoFake. All the speech data converted through emotion voice conversion constitutes the fake audio of EmoFake.

We choose the emotion fake audio generated by S1 and S2 as fake audio and emotional speech corresponding to the speaker in ESD as real audio of the train and dev set. In order to evaluate generalization of the models, we made the test set unseen. We choose S3, S4, S5, S6 and S7 to generate fake audio of the test set. The real audio is also from the emotional speech of ESD. The specific information of EmoFake is shown in Table 3.

## 3 EXPERIMENTS

### 3.1 Experiments setup

We chose five fake audio detection models commonly used as baseline models in most fake audio detection systems for evaluation: Light Convolutional Neural Network (LCNN) (Wang and Yamagishi, 2021), RawNet2 (Tak et al., 2021), SENet (Hu et al., 2018), ResNet34 (He et al., 2016) and AASIST (Jung et al., 2022).

Linear frequency cepstral coefficients (LFCC) (Sahidullah et al., 2015) is used as input feature to LCNN[9], SENet[10] and ResNet34. LFCC are extracted using a 20 ms window with a 10 ms shift, a 1024-point Fourier transform and 70 filters. The LCNN model consists of five convolution layers, ten Max-Feature-Map layers and two fully connected (FC) layers. The more detailed information about the LCNN model can be found in (Wang and Yamagishi, 2021). As a fully end-to-end system, RawNet2[9] uses a fixed bank of sinc filters to operate raw audio waveforms. Then it uses six residual blocks which is followed by a gated recurrent units (GRU) and FC layers to produce two-class predictions: bonafide or spoof. The more detailed information about the RawNet2 we used can be found in (Tak et al., 2021). The SENet we utilized is composed of three SEBlocks. Each SEBlock comprises two convolutional layers and one SELayer (Hu et al., 2018). The framework of ResNet34 we used is the same as (He et al., 2016). AASIST[11] is an end-to-end audio anti-spoofing system that employs integrated spectro-temporal graph attention networks. It introduces a novel heterogeneous stacking graph attention layer and a max

[9]https://github.com/asvspoof-challenge/2021
[10]https://github.com/moskomule/senet.pytorch
[11]https://github.com/clovaai/aasist

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1286-1297, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Table 3: The description of the EmoFake dataset composition.

| Subset | Part | Speaker pair | Source emotion | Target emotion | Fake | | | | | | | | Real |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | #S1 | #S2 | #S3 | #S4 | #S5 | #S6 | #S7 | #Total | #Total |
| English | Train | E1 | Neu | Ang | 700 | 700 | 0 | 0 | 0 | 0 | 0 | 5,600 | 3,500 |
| | | | | Hap | 700 | 700 | 0 | 0 | 0 | 0 | 0 | | |
| | | | | Sad | 700 | 700 | 0 | 0 | 0 | 0 | 0 | | |
| | | | | Sur | 700 | 700 | 0 | 0 | 0 | 0 | 0 | | |
| | | E2 | Ang | Neu | 700 | 700 | 0 | 0 | 0 | 0 | 0 | 5,600 | 3,500 |
| | | | | Hap | 700 | 700 | 0 | 0 | 0 | 0 | 0 | | |
| | | | | Sad | 700 | 700 | 0 | 0 | 0 | 0 | 0 | | |
| | | | | Sur | 700 | 700 | 0 | 0 | 0 | 0 | 0 | | |
| | | E3 | Hap | Neu | 700 | 700 | 0 | 0 | 0 | 0 | 0 | 5,600 | 3,500 |
| | | | | Ang | 700 | 700 | 0 | 0 | 0 | 0 | 0 | | |
| | | | | Sad | 700 | 700 | 0 | 0 | 0 | 0 | 0 | | |
| | | | | Sur | 700 | 700 | 0 | 0 | 0 | 0 | 0 | | |
| | Dev | E4 | Sur | Neu | 700 | 700 | 0 | 0 | 0 | 0 | 0 | 5,600 | 3,500 |
| | | | | Ang | 700 | 700 | 0 | 0 | 0 | 0 | 0 | | |
| | | | | Hap | 700 | 700 | 0 | 0 | 0 | 0 | 0 | | |
| | | | | Sad | 700 | 700 | 0 | 0 | 0 | 0 | 0 | | |
| | Test | E5 | Sad | Neu | 0 | 0 | 700 | 700 | 700 | 700 | 700 | 14,000 | 3,500 |
| | | | | Ang | 0 | 0 | 700 | 700 | 700 | 700 | 700 | | |
| | | | | Hap | 0 | 0 | 700 | 700 | 700 | 700 | 700 | | |
| | | | | Sur | 0 | 0 | 700 | 700 | 700 | 700 | 700 | | |
| Chinese | Train | C1 | Neu | Ang | 700 | 700 | 0 | 0 | 0 | 0 | 0 | 5,600 | 3,500 |
| | | | | Hap | 700 | 700 | 0 | 0 | 0 | 0 | 0 | | |
| | | | | Sad | 700 | 700 | 0 | 0 | 0 | 0 | 0 | | |
| | | | | Sur | 700 | 700 | 0 | 0 | 0 | 0 | 0 | | |
| | | C2 | Ang | Neu | 700 | 700 | 0 | 0 | 0 | 0 | 0 | 5,600 | 3,500 |
| | | | | Hap | 700 | 700 | 0 | 0 | 0 | 0 | 0 | | |
| | | | | Sad | 700 | 700 | 0 | 0 | 0 | 0 | 0 | | |
| | | | | Sur | 700 | 700 | 0 | 0 | 0 | 0 | 0 | | |
| | | C3 | Hap | Neu | 700 | 700 | 0 | 0 | 0 | 0 | 0 | 5,600 | 3,500 |
| | | | | Ang | 700 | 700 | 0 | 0 | 0 | 0 | 0 | | |
| | | | | Sad | 700 | 700 | 0 | 0 | 0 | 0 | 0 | | |
| | | | | Sur | 700 | 700 | 0 | 0 | 0 | 0 | 0 | | |
| | Dev | C4 | Sur | Neu | 700 | 700 | 0 | 0 | 0 | 0 | 0 | 5,600 | 3,500 |
| | | | | Ang | 700 | 700 | 0 | 0 | 0 | 0 | 0 | | |
| | | | | Hap | 700 | 700 | 0 | 0 | 0 | 0 | 0 | | |
| | | | | Sad | 700 | 700 | 0 | 0 | 0 | 0 | 0 | | |
| | Test | C5 | Sad | Neu | 0 | 0 | 700 | 700 | 700 | 700 | 700 | 14,000 | 3,500 |
| | | | | Ang | 0 | 0 | 700 | 700 | 700 | 700 | 700 | | |
| | | | | Hap | 0 | 0 | 700 | 700 | 700 | 700 | 700 | | |
| | | | | Sur | 0 | 0 | 700 | 700 | 700 | 700 | 700 | | |

graph operation to model artifacts across different temporal and spectral intervals. Specific information about AASIST can be obtained in (Jung et al., 2022). All models are implemented using the pytorch framework and trained on a single NVIDIA 2080ti. To ensure the reliability of the experimental results, all our results are the average of three trials.

### 3.2 Evaluation metrics

For speech emotion recognition experiments, we use accuracy rate (ACC) as the evaluation metrics. ACC is defined as the ratio of the number of samples with correctly predicted labels after recognition by the

CCL 2024

speech emotion recognition model to the total number of samples. It can be expressed as:

$$\mathcal{ACC} = \frac{\sum_{i=1}^{n} \phi(y_i, f(x_i))}{n} \tag{1}$$

$y_i$ is the real label corresponding to sample $x_i$, $n$ is the total number of samples, $f(x_i)$ is the corresponding predicted label generated by sample $x_i$ through the speech emotion recognition model $f$, and the function $\phi$ is defined as follows:

$$\phi(x, y) = \begin{cases} 1 & x = y \\ 0 & x \neq y \end{cases} \tag{2}$$

In addition, in multi-classification problems, for cases where samples are imbalanced, weighted accuracy (WACC) is an accuracy calculation method that considers the number of samples in each category. It is calculated by multiplying the accuracy of each category by the proportion of that category in the total sample and then summing it. The calculation is as follows:

$$\text{WACC} = \frac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N} TP_i + FP_i} \tag{3}$$

$N$ represents the total number of categories, $TP$ represents the number of samples that are both predicted and actually positive, and $FP$ represents the number of samples that are predicted to be positive but actually are negative.

In order to evaluate different fake audio detection models, the equal error rate (EER) (Wu et al., 2015b) is used as the evaluation metrics. Let $P_{\text{fa}}(\theta)$ and $P_{\text{miss}}(\theta)$ be the false alarm and miss rates at threshold $\theta$ defined according to:

$$P_{\text{fa}}(\theta) = \frac{\#\{\text{spoofed trials with score} > \theta\}}{\#\{\text{total spoofed trials}\}} \tag{4}$$

$$P_{\text{miss}}(\theta) = \frac{\#\{\text{genuine trials with score } \leq \theta\}}{\#\{\text{total genuine trials}\}} \tag{5}$$

$$EER = P_{\text{miss}}(\theta) = P_{\text{miss}}(\theta) \tag{6}$$

$P_{\text{fa}}(\theta)$ and $P_{\text{miss}}(\theta)$ are, respectively, monotonically decreasing and increasing functions of $\theta$. As shown in equation(5), the EER corresponds to the threshold $\theta_{\text{EER}}$ at which the two detection error rates are equal. The higher the EER, the more difficult it is for the model to discriminate the fake audio.

### 3.3 Performance of speech emotion recognition

To evaluate the performance of emotion fake audio, we conducted emotion recognition experiment. The hierarchical grained and feature model (HGFM)[12] (Xu et al., 2020) was chosen as the speech emotion recognition model. It made further improvements based on the hierarchical GRU network (Jiao et al., 2019). The frame-level module of HGFM learns frame information through bi-directional GRU model, and the utterancelevel module uses frame-level structure outputs fused with statistical features to learn the final utterance representation.

In order to eliminate the influence of the language environment on HGFM, the ESD data set is divided into an English subset (speakers correspond to 0011 to 0020) and a Chinese subset (speakers correspond to 0001 to 0010). Then we use the English subset of ESD, all the fake audio in the English subset of EmoFake, the Chinese subset of ESD and all the fake audio in the Chinese subset of EmoFake to train four HGFM models. Then the trained model is used for speech emotion recognition experiments on the data used for training. For the recognition accuracy of each emotion, we use ACC as the evaluation metric. For the overall recognition accuracy, WACC is used as the evaluation metric. The experiment results are shown in Table 4. For both two languages, the WACC of EmoFake is lower than ESD. However, the WACC for emotion recognition in two languages with emotion forgery speech still exceeds 90%, and

CCL 2024

Table 4: Results of speech emotion recognition experiments.

| Language | Dataset | Emotion (ACC(%)) | | | | | Total (WACC(%)) |
|---|---|---|---|---|---|---|---|
| | | Neu | Hap | Sad | Ang | Sur | |
| English | ESD (0011-0020) | 98.86 | 97.37 | 98.43 | 98.31 | 97.26 | 98.05 |
| | EmoFake English subset | 96.09 | 87.12 | 90.30 | 88.29 | 92.48 | 90.89 |
| Chinese | ESD (0001-0010) | 99.69 | 98.80 | 99.86 | 99.91 | 99.00 | 99.45 |
| | EmoFake Chinese subset | 93.73 | 95.58 | 92.09 | 93.32 | 90.09 | 93.01 |

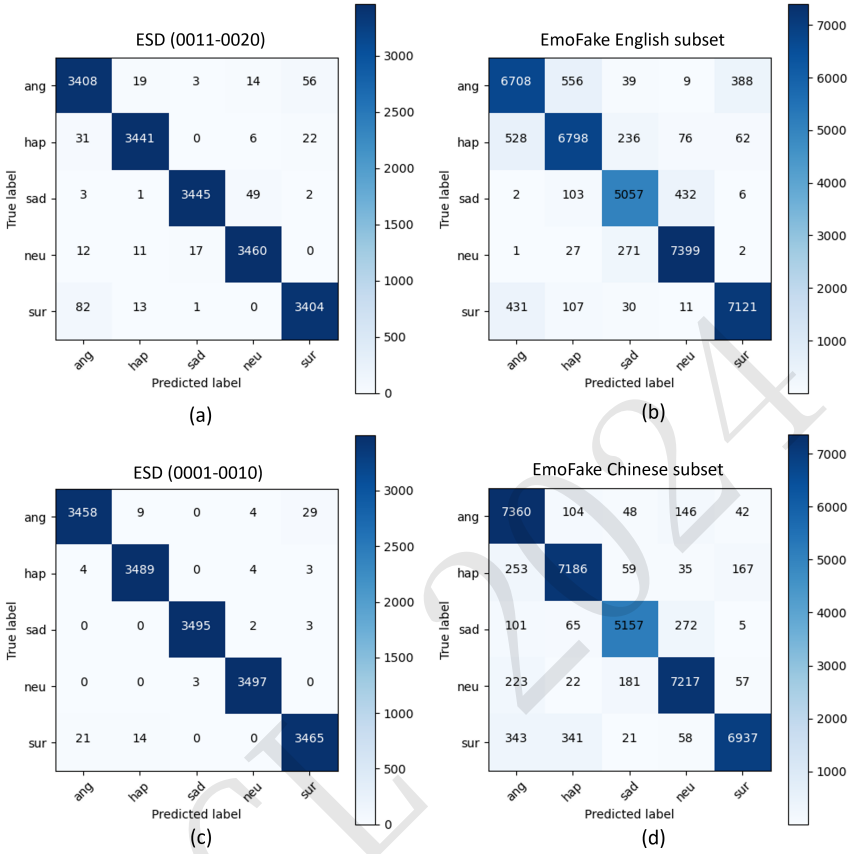



Figure 2: The confusion matrix of emotion recognition experiments.(a) is the confusion matrix tested on the English subset of ESD. (b) is the confusion matrix tested on the English subset of EmoFake. (c) is the confusion matrix tested on the Chinese subset of ESD. (d) is the confusion matrix tested on the Chinese subset of EmoFake.

the model performs better in Chinese emotion fake audio. The results indicate that while not as high as ESD, EmoFake still exhibits a relatively high level of emotional consistency.

To observe the emotional consistency of EmoFake more intuitively, confusion matrixs of speech emotion recognition experiments are provided as shown in Figure 3. Compared with the confusion matrix obtained from real emotional speech, the diagonal colors in the confusion matrix obtained from the emotion recognition experiment using EmoFake are lighter, indicating that the number of correctly recognized speech samples by HGFM is relatively small.

In addition, we conducted experiment on emotion fake audio generated by different EVC models. The experiment results are shown in Table 5. For S3, S4, S5, S6 and S7, we only used them to convert Sad speech to other four target emotions. Therefore, there are no results for the Sad emotion in Table 5 for these models. From the experiment results, it can be observed that there are significant differences among

[12]https://github.com/zjyyyy/HGFM

Table 5: Results of evaluating speech emotion recognition model using emotion fake audio generated by different EVC models. ”-” stands for no fake audio at this location.

| EVC model | Target emotion(ACC (%)) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | English subset | | | | | Chinese subset | | | | |
| | Neu | Hap | Sad | Ang | Sur | Neu | Hap | Sad | Ang | Sur |
| S1 | 98.76 | 96.67 | 98.71 | 97.00 | 95.67 | 99.52 | 98.14 | 99.61 | 96.57 | 92.86 |
| S2 | 70.62 | 62.57 | 76.29 | 60.76 | 55.33 | 84.14 | 81.00 | 68.71 | 65.29 | 49.38 |
| S3 | 98.29 | 98.86 | - | 97.71 | 98.29 | 54.00 | 59.14 | - | 71.00 | 69.14 |
| S4 | 98.43 | 96.43 | - | 97.71 | 95.86 | 99.14 | 98.86 | - | 98.71 | 91.29 |
| S5 | 99.57 | 41.71 | - | 57.43 | 98.57 | 100.00 | 99.14 | - | 98.86 | 93.29 |
| S6 | 97.43 | 94.43 | - | 97.29 | 96.00 | 99.71 | 98.57 | - | 98.43 | 92.29 |
| S7 | 99.71 | 100.00 | - | 99.43 | 99.57 | 99.86 | 99.71 | - | 99.86 | 99.29 |

Table 6: Results of experiment on fake audio detection models trained with the train and dev subset of the LA of ASVspoof 2019 dataset. Evaluation metric is EER (%)

| Fake audio detection model | ASVspoof 2019 eval | ASVspoof 2021 eval | EmoFake English subset test | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Total | S3 | S4 | S5 | S6 | S7 |
| LCNN | 7.62 | 13.45 | 13.93 | **16.33** | 10.94 | 17.83 | 10.77 | 7.82 |
| RawNet2 | 5.69 | **13.14** | 26.54 | 28.00 | 21.20 | 45.56 | 20.36 | 5.38 |
| SENet | 6.02 | 22.50 | 45.03 | 38.71 | 49.43 | 65.09 | 49.34 | 18.95 |
| ResNet34 | 3.06 | 23.21 | 25.21 | 20.72 | 24.10 | 40.05 | 23.37 | 13.95 |
| AASIST | **1.96** | 16.11 | **9.83** | 26.40 | **4.00** | **6.40** | **4.19** | **1.45** |

different EVC models in handling emotion voice conversion tasks. We believe that the high recognition accuracy of speech emotion recognition experiments represents the completion of emotion conversion in emotion fake audio. In the English experiment, emotion fake audio generated by S2 had the worst conversion completion of emotions. In the Chinese experiment, there was not much difference in the conversion completion of emotion fake audio generated by S2 and S3. Some models (such as S1 and S7) can accurately convert the target emotion. These differences may be related to the design of the models, the diversity of training data, and the ability of the models to capture different emotional features.

### 3.4 Performance of fake audio detection models

To verify the threat of emotion fake audio to existing fake audio detection models, we evaluated some commonly used fake audio detection models using EmoFake. To eliminate the influence of language on model discrimination ability, we trained the fake audio detection model using both the Chinese fake auido dataset and the English fake audio dataset. We used the LA dataset of ASVspoof 2019 as the English dataset to train all English fake audio detection models. Then we choose the evaluation set of LA dataset of ASVspoof 2019, the evaluation set of LA dataset of ASVspoof 2021 and the test part of the Enlgish subset of EmoFake to evaluate fake audio detection models. The results are shown in Table 6. As Chinese evaluation, the track 3.2 of ADD 2022 dataset was chosen to train all Chinese fake audio detection models. The test set of round 1 of track 3.2 of ADD 2022 dataset, the test set of round 1 of track 1.2 of ADD 2023 dataset and the test part of Chinese subset of EmoFake were used to evaluated Chinese fake audio detection models. The results are shown in Table 7.

When evaluating the English fake audio detection models using the English subset of EmoFake, compared to the results of the eval of ASVspoof 2019 and 2021, all models exhibited an increase in EER. This indicates that emotion fake audio has a detrimental effect on the performance of fake audio detection models. The performance degradation of SENet and RawNet2 is the most severe. And SENet's EER increased by 16.48 when evaluating using ASVspoof 2021 eval compared to evaluating using ASVspoof 2021 eval. It can be seen that SENet has poor robustness when facing unknown fake audio. Among

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1286-1297, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Table 7: Results of experiment on fake audio detection models trained with the train and dev subset of the track 3.2 of ADD 2022 dataset. Evaluation metric is EER (%)

| Fake audio detection model | ADD 2022 track 3.2 R1 | ADD 2023 track 1.2 R1 | EmoFake Chinese subset test | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Total | S3 | S4 | S5 | S6 | S7 |
| LCNN | 34.01 | 58.46 | 29.69 | 26.29 | 23.97 | 4.65 | 30.29 | 54.44 |
| RawNet2 | 52.38 | 56.78 | 38.85 | 21.40 | 46.15 | 17.88 | 51.19 | 52.79 |
| SENet | 57.71 | 54.13 | 31.97 | 35.30 | 25.90 | 16.69 | 31.26 | 46.09 |
| ResNet34 | 53.31 | 52.08 | 34.50 | 38.06 | 35.28 | 19.07 | 37.08 | 44.27 |
| AASIST | **26.85** | **46.73** | **9.50** | **3.14** | **9.91** | **0.76** | **14.06** | **13.17** |

the five fake audio detection models we selected, AASIST performed the best when evaluated using the English subset of EmoFake. However, compared to the results on ASVspoof 2019, it still improved by 7.87.

The ADD 2022 and ADD 2023 datasets have added many new types of fake audio, such as partially fake audio and scene fake audio. Multiple types of fake audio poses greater challenges to the performance of fake audio detection models. EmoFake only includes real emotional speech and one type of fake audio which is emotion fake audio. Therefore, the EER evaluated using the ADD 2022 and ADD 2023 datasets for the five Chinese fake audio detection models is higher than that evaluated using the Chinese subset of EmoFake. However, the performance degradation brought by emotion fake audio to Chinese fake audio detection models cannot be ignored. Similar to the results obtained from evaluating using English datasets, AASIST performed the most robust among the five models.

In order to evaluate which EVC model generated more undetectable fake audio, we evaluated fake audio detection models using the emotion fake audio in the test set of EmoFake generated by S3, S4, S5, S6 and S7 separately. An interesting phenomenon is that for English fake audio detection models, emotion fake audio generated by S5 is the most difficult to discriminate, while emotion fake audio generated by S7 is the easiest. The performance of the Chinese fake audio detection models is exactly the opposite. The emotion fake audio generated by S7 is the most difficult to discriminate, while the emotional fake audio generated by S5 is the easiest. The reason for this result may be related to whether the structure of the fake audio generation model in the data used to train the fake audio detection models is similar to some EVC models structures in EmoFake.

### 3.5 Baselines trained with EmoFake dataset

To provide the baseline fake audio detection models trained with emotion fake audio, we retrained five models with the English subset and Chinese subset of EmoFake separately. Next, we evaluated the baseline models trained with emotion fake audio using the test part of EmoFake. The evaluation results of the English baseline models are shown in Table 8, and the evaluation results of the Chinese baseline models are shown in Table 9. The experiment results illustrate that most baseline models have enhanced their capability to discriminate emotion fake audio after training with EmoFake. AASIST performs the best overall in the English baseline models. The EER evaluated using EmoFake by AASIST trained with emotion fake audio was 6.18 lower than that evaluated using EmoFake by AASIST trained with ASVspoof 2019. The best performing Chinese baseline model is RawNet2. Except for AASIST, all Chinese fake audio detection models have improved performance after training with emotion fake audio. Comparing the evaluation results of AASIST in Tables 9 and 7, it can be found that AASIST trained with EmoFake only shows a decrease in discriminative ability for emotion fake audio generated by S3. We speculate that the training data of ADD 2022 may include fake audio generated by models similar to the Seq2Seq structure used in S3, while the training data in EmoFake does not. This proves that our strategy of generating emotion fake audio in the test part with EVC models that do not overlap with the model structure in the train and dev part has a certain effect.

Additionally, we evaluated our baseline models using emotion fake audio with different emotions in the test part of EmoFake. Notably, because the source emotion in the test part is Sad, the recorded results

CCL 2024

Table 8: Results of experiment on fake audio detection models trained with the train and dev part of the English subset of EmoFake.

| Fake audio detection model | Total | EmoFake English subset test (EER(%)) EVC model | | | | | Target emotion | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | S3 | S4 | S5 | S6 | S7 | Neu | Hap | Sur | Ang |
| LCNN | 6.16 | 10.60 | 9.98 | 7.10 | 5.04 | 2.73 | 7.10 | 6.33 | 4.25 | **7.21** |
| RawNet2 | 7.92 | 16.80 | 1.53 | 0.01 | 1.30 | 6.97 | 5.51 | 4.06 | 7.26 | 7.22 |
| SENet | 18.25 | 49.12 | 3.52 | 3.08 | 4.76 | 7.93 | 18.31 | 19.27 | 17.26 | 18.01 |
| ResNet34 | 16.90 | 46.37 | 2.58 | 2.25 | 3.13 | 3.13 | 16.01 | 17.60 | 17.52 | 17.57 |
| AASIST | **3.65** | **8.65** | **0.01** | **0.00** | **0.00** | **1.31** | **1.23** | **3.98** | **3.80** | 17.57 |

Table 9: Results of experiment on fake audio detection models trained with the train and dev part of the Chinese part of EmoFake.

| Fake audio detection model | Total | EmoFake Chinese subset test (EER(%)) EVC model | | | | | Target emotion | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | S3 | S4 | S5 | S6 | S7 | Neu | Hap | Sur | Ang |
| LCNN | 9.24 | **11.21** | 3.24 | **0.53** | 8.43 | 16.85 | 10.75 | **7.69** | 14.72 | **4.70** |
| RawNet2 | **8.34** | 13.65 | 12.76 | 1.28 | 1.86 | 4.25 | **4.51** | 10.14 | 11.44 | 7.86 |
| SENet | 15.95 | 45.37 | 0.96 | 0.73 | 1.14 | **3.96** | 14.72 | 15.48 | 16.81 | 15.18 |
| ResNet34 | 15.51 | 38.24 | 0.91 | 0.94 | 1.01 | 12.68 | 15.31 | 15.74 | 15.72 | 15.96 |
| AASIST | 11.32 | 23.20 | **0.41** | 6.45 | **0.43** | 5.67 | 15.43 | 11.53 | **10.35** | 8.59 |

do not include data for the Sad emotion. For different fake audio detection models, the emotion detecting difficultly is also different.

## 4 CONCLUSION

This work describes the design strategy of the EmoFake dataset, including the generation of fake data and the selection of real data. As far as we know, EmoFake is the first dataset designed for detecting the emotion fake audio. The emotion fake audio in EmoFake is generated using seven open source emotion voice conversion models. This work also evaluates the performance of existing fake audio detection models facing emotion fake audio. The results show that the fake audio detection model trained with the LA dataset of ASVspoof 2019 and the track 3.2 of ADD 2022 dataset does have a reduced discriminative ability for emotion fake audio. We believe that the EmoFake dataset will further accelerate research in fake audio detection and inspires broader advancements in the field. Future work involves adding more emotion voice conversion models and expanding EmoFake to include multiple languages. Furthermore, we will explore how varying emotional intensities impact fake audio detection while maintaining a consistent emotional category.

## Acknowledgements

This work is supported by the National Natural Science Foundation of China (NSFC) (No. 62322120, No. 62306316, No.U21B2010, No. 62206278).

## References

Mingyi Chen, Xuanji He, Jing Yang, et al. 2018. 3-d convolutional recurrent neural networks with attention model for speech emotion recognition. *IEEE Signal Processing Letters*.

Changzeng Fu, Chaoran Liu, Carlos Toshinori Ishi, and Hiroshi Ishiguro. 2022. An improved cyclegan-based emotional voice conversion model by augmenting temporal dependency with a transformer. *Speech Communication*, 144:110–121.

Jian Gao, Deep Chakraborty, Hamidou Tembine, and Olaitan Olaleye. 2019. Nonparallel Emotional Speech Conversion. In *Proc. Interspeech 2019*, pages 2858–2862.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*, June.

Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June.

Wenxiang Jiao, Haiqin Yang, Irwin King, and others. 2019. HiGRU: Hierarchical gated recurrent units for utterance-level emotion recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 397–406.

Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, et al. 2022. Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *ICASSP 2022*, pages 6367–6371.

Tomi Kinnunen, Md. Sahidullah, Héctor Delgado, et al. 2017. The ASVspoof 2017 Challenge: Assessing the Limits of Replay Spoofing Attack Detection. In *Proc. Interspeech 2017*, pages 2–6.

Nicolas M Müller, Piotr Kawa, Wei Herng Choong, Edresson Casanova, Eren Gölge, Thorsten Müller, Piotr Syga, Philip Sperl, and Konstantin Böttinger. 2024. Mlaad: The multi-language audio anti-spoofing dataset. *arXiv preprint arXiv:2401.09512*.

Georgios Rizos, Alice Baird, et al. 2020. Stargan for emotional speech conversion: Validated by data augmentation of end-to-end emotion recognition. In *ICASSP 2020*, pages 3502–3506.

Md. Sahidullah, Tomi Kinnunen, and Cemal Hanilçi. 2015. A comparison of features for synthetic speech detection. In *Proc. Interspeech 2015*, pages 2087–2091.

Hemlata Tak, Jose Patino, Massimiliano Todisco, et al. 2021. End-to-end anti-spoofing with rawnet2. In *ICASSP 2021*, pages 6369–6373.

Massimiliano Todisco, Xin Wang, Ville Vestman, et al. 2019. ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection. In *Proc. Interspeech 2019*, pages 1008–1012.

Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Xin Wang and Junichi Yamagishi. 2021. A Comparative Study on Recent Neural Spoofing Countermeasures for Synthetic Speech Detection. In *Proc. Interspeech 2021*, pages 4259–4263.

Zhizheng Wu, Ali Khodabakhsh, Cenk Demiroglu, et al. 2015a. Sas: A speaker verification spoofing database containing diverse attacks. In *ICASSP 2015*, pages 4440–4444.

Zhizheng Wu, Tomi Kinnunen, Nicholas Evans, et al. 2015b. ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge. In *Proc. Interspeech 2015*, pages 2037–2041.

Yunfeng Xu, Hua Xu, and Jiyun Zou. 2020. Hgfm: A hierarchical grained and feature model for acoustic emotion recognition. In *ICASSP 2020*, pages 6499–6503. IEEE.

Junichi Yamagishi, Xin Wang, Massimiliano Todisco, et al. 2021. ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection. In *Proc. 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, pages 47–54.

Jiangyan Yi, Ye Bai, Jianhua Tao, Haoxin Ma, et al. 2021. Half-Truth: A Partially Fake Audio Detection Dataset. In *Proc. Interspeech 2021*, pages 1654–1658.

Jiangyan Yi, Ruibo Fu, Jianhua Tao, et al. 2022. Add 2022: the first audio deep synthesis detection challenge. In *ICASSP 2022*, pages 9216–9220.

Jiangyan Yi, Jianhua Tao, Ruibo Fu, et al. 2023a. Add 2023: the second audio deepfake detection challenge. *arXiv preprint arXiv:2305.13774*.

Jiangyan Yi, Chenglong Wang, Jianhua Tao, Xiaohui Zhang, Chu Yuan Zhang, and Yan Zhao. 2023b. Audio deepfake detection: A survey. *arXiv preprint arXiv:2308.14970*.

CCL 2024

Jiangyan Yi, Chenglong Wang, Jianhua Tao, Chu Yuan Zhang, Cunhang Fan, Zhengkun Tian, Haoxin Ma, and Ruibo Fu. 2024. Scenefake: An initial dataset and benchmarks for scene fake audio detection. *Pattern Recognition*, 152:110468.

Yongyi Zang, You Zhang, Mojtaba Heydari, and Zhiyao Duan. 2024. Singfake: Singing voice deepfake detection. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12156–12160.

Kun Zhou, Berrak Sisman, and Haizhou Li. 2020a. Transforming Spectrum and Prosody for Emotional Voice Conversion with Non-Parallel Training Data. In *Proc. The Speaker and Language Recognition Workshop (Odyssey 2020)*, pages 230–237.

Kun Zhou, Berrak Sisman, Mingyang Zhang, and Haizhou Li. 2020b. Converting Anyone's Emotion: Towards Speaker-Independent Emotional Voice Conversion. In *Proc. Interspeech 2020*, pages 3416–3420.

Kun Zhou, Berrak Sisman, and Haizhou Li. 2021a. Limited Data Emotional Voice Conversion Leveraging Text-to-Speech: Two-Stage Sequence-to-Sequence Training. In *Proc. Interspeech 2021*, pages 811–815.

Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. 2021b. Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset. In *ICASSP 2021*, pages 920–924.

Kun Zhou, Berrak Sisman, et al. 2022. Emotional voice conversion: Theory, databases and esd. *Speech Communication*.

CCL 2024

# Going Beyond Passages: Readability Assessment for Book-level Long Texts

**Wenbiao Li**[1,2], **Rui Sun**[1,2], **Tianyi Zhang**[1,2], **Yunfang Wu** [1,3*]
[1]MOE Key Laboratory of Computational Linguistics, Peking University
[2]School of Software & Microelectronics, Peking University
[3]School of Computer Science, Peking University, Beijing, China
{2001210322,2101210366,2301210557}@stu.pku.edu.cn, wuyf@pku.edu.cn

## Abstract

Readability assessment for book-level long text is widely needed in real educational applications. However, most of the current researches focus on passage-level readability assessment and little work has been done to process ultra-long texts. In order to process the long sequence of book texts better and to enhance pretrained models with difficulty knowledge, we propose a novel model DSDR, difficulty-aware segment pre-training and difficulty multi-view representation. Specifically, we split all books into multiple fixed-length segments and employ unsupervised clustering to obtain difficulty-aware segments, which are used to re-train the pretrained model to learn difficulty knowledge. Accordingly, a long text is represented by averaging multiple vectors of segments with varying difficulty levels. We construct a new dataset of Graded Children's Books to evaluate model performance. Our proposed model achieves promising results, outperforming both the traditional SVM classifier and several popular pretrained models. In addition, our work establishes a new prototype for book-level readability assessment, which provides an important benchmark for related research in future work.

## 1 Introduction

The readability assessment is to automatically predict the difficulty level of an input text, which has a wide range of applications, such as automating reader advisory (Pera and Ng, 2014), clinical informed consent forms (Perni et al., 2019) and patient education based on the Internet (Sare et al., 2020).

The readability assessment has a long history of research. In early days, researchers focus on scoring the reading difficulty of texts by designing various readability formulas. Later, many works treat it as a classification task, and various linguistic features are exploited to perform machine learning classification (Schwarm and Ostendorf, 2005; Petersen and Ostendorf, 2009; Qiu et al., 2017). In recent years, traditional linguistic features are combined with the pre-trained model to predict the difficulty level of a text (Qiu et al., 2021; Li et al., 2022). In addition, there are works regarding readability assessment as a regression task (Sheehan et al., 2010) or a ranking task (Ma et al., 2012; Lee and Vajjala, 2022).

Today, reading a whole book plays a vital role to improve children's reading comprehension ability, thus followed by an increasing demand for the book readability assessment. However, most of the existing work processes readability assessment with a passage as input, which is usually an extracted part of a book. These passage-level models have limitations when it comes to the assessment of book-level long texts.

Going from passage-level to book-level readability assessment, a challenging task is how to deal with the long input sequence. As is illustrated in Figure 1, the length of a book text is typically tens times more than that of a passage text. To handle the long sequence simultaneously is a hard work. Taking the pre-trained model BERT as an example, when the mini-batch is 1, using a 24G video memory GPU, the maximum length of slices should be 16 * 512 = 8192. In our children's books data, only 23% [74/320] of the samples were below this value in length. Moreover, the widely applied strategies to process long

*Corresponding author.
©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1298-1309, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers

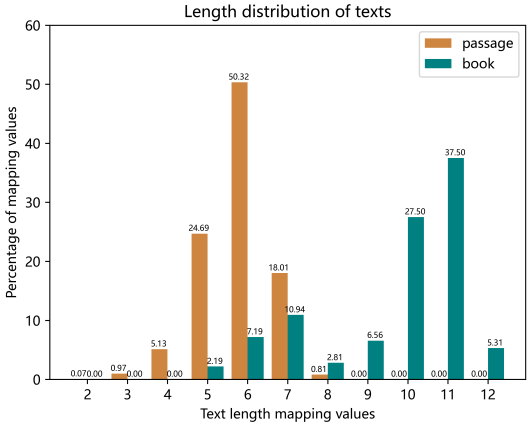



Figure 1: Statistics of length distribution comparing passages and books, derived from textbooks and children's books respectively. Let the text length be $x$, the mapping function is $f(x) = \lfloor ln(x) \rfloor$.

sequence in passage-level text applications, such as truncating, sliding window and pooling, could not work well for book-level long texts because of too much loss of information.

Nowadays, pre-trained models have made remarkable achievements in many NLP tasks. But these models, such as the pioneer BERT (Devlin et al., 2018) and the long-sequence-specialized BigBird (Zaheer et al., 2020), are mostly based on masked language modeling and next sentence prediction, and thus can not express difficulty knowledge. Moreover, the computational complexity of the multi-head self-attention mechanism in the model structure is proportional to the square of the length of the input sequence. Consequently, directly applying pre-trained models to book readability assessment cannot obtain satisfying results.

In this paper, Firstly, we investigate whether the passage-level readability assessment model can be applied to predict the reading difficulty of a book, and we get a negative answer, indicating that there exists a large gap between passage-level and book-level readability assessment.

Therefore, we propose a two-stage model with **d**ifficulty-aware **s**egment pre-training and multi-view **d**ifficulty **r**epresentation (DSDR) for this challenging task. Specifically, we split all book texts into fixed-length segments, and perform unsupervised clustering to assign each segment with a difficulty label according to linguistic features. The labeled segment difficulty data is exploited to finetune the pre-trained model with a difficulty classification task, in order to enhance the pre-trained model with the ability to recognize and represent difficulty knowledge. In the next stage, the difficulty-aware pre-trained model is employed to extract difficulty features of segments, and multiple difficulty vectors are fused to form the overall representation of a whole book.

We construct a Graded Children's Books (GCB) for book-level readability assessment, including 320 popular stories and famous novels in Chinese. We conduct detailed experiments on this dataset. The experimental results demonstrate that our proposed model achieves an accuracy of 73.125, which significantly outperforms the traditional machine learning method SVM, neural network classifiers and pretrained models. We will make our data and code publicly available for future research.

To sum up, our contributions are as follows:

- We address the readability assessment of book-level long text with pre-trained language models, and construct a Chinese dataset of children's books labelled with difficulty levels.
- We propose a novel DSDR model to deal with the long sequence effectively and enrich the pre-trained models with difficulty knowledge.
- Our proposed model achieves promising results for long text readability assessment, outperforming both the traditional classifier and popular pre-trained models.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1298–1309, Taiyuan, China, July 25 – 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

## 2 Related Work

### 2.1 Readability Assessment

In the early days, researchers developed a variety of readability formulas, including Flesch (Flesch, 1948), Dale-Chall (Dale and Chall, 1948) and SMOG (Mc Laughlin, 1969). In the field of NLP, readability assessment is often regarded as a classification problem.

For traditional machine learning methods, researchers have used various linguistic features defined by linguists to construct models (Schwarm and Ostendorf, 2005; Petersen and Ostendorf, 2009; Qiu et al., 2017; Lu et al., 2019). Some works (Imperial, 2021; Deutsch et al., 2020; Lee et al., 2021) combined features extracted from neural networks with linguistic features and use traditional machine learning algorithms to classify.

For deep learning methods, (Jiang et al., 2018) purposes knowledge-enriched word embedding (KEWE), which encodes the knowledge of reading difficulty into the representation of words. (Azpiazu and Pera, 2019) present a multi-attentive recurrent neural network architecture for automatic multilingual readability assessment. There are also some works (Qiu et al., 2021; Li et al., 2022) that fuse neural network features with artificial linguistic features.

Besides English and Chinese texts, (Reyes et al., 2022) develop a readability assessment model for Cebuano. However, the existing works don't specialize for long text.

### 2.2 Long Sequence Processing

For book-level readability assessment, the process of the long sequence input is a challenging task because of the excessive computational complexity of full attention in long sequence processing. Current work on reducing the computational complexity exploits the following methods.

**Sparse attention.** It reduces the computational complexity by keeping only the connections of surrounding tokens (Beltagy et al., 2020; Zaheer et al., 2020).

**Approximation methods.** Reformer (Kitaev et al., 2020) replaces dot-product attention by one that uses locality-sensitive hashing and uses reversible residual layers instead of the standard residuals. Performers (Choromanski et al., 2020) use a Fast Attention Via positive Orthogonal Random features approach (FAVOR+) to approximate softmax attention kernels. Linformer (Wang et al., 2020) demonstrates that the self-attention mechanism can be approximated by a low-rank matrix.

**Alternative methods.** FNet (Lee-Thorp et al., 2021) replaces the self-attention sublayers with simple linear transformations PoNet (Tan et al., 2021) uses a Pooling Network for token mixing in long sequences with a linear complexity.

**Compression methods.** Pyraformer (Liu et al., 2021) introduces the pyramidal attention module (PAM). Triformer (Cirstea et al., 2022) uses a triangular, variable-specific attention which contains Linear complexity and Variable-specific parameters.

In our work, we choose BigBird with sparse attention as a baseline model and borrow the compression methods to implement PyramidFormer.

### 2.3 Multi-representation Modeling

In the task of pairwise comparison between sequences, there are usually two strategies, Cross-encoder and Bi-encoder, the former works well but is too slow. Poly-encoder (Humeau et al., 2019) is proposed as a compromise between the two methods. Drawing on the idea of Poly-encoder, some multi-representation work was born. (Kong et al., 2022) design a lightweight network to fuse the aspect embeddings for representing queries and documents. (Xu et al., 2022) introduce a novel multi-task model called Mixture of Virtual-Kernel Experts (MVKE) to learn user preferences on various actions and topics unitedly.

## 3 Can Passage-level Readability Assessment Model Apply to Long Texts?

Comparing with passage-level readability assessment, book-level readability assessment is less addressed and more challenging. For both levels, our experimental datasets come from the educational

CCL 2024

field, and these texts are selected by experts according to children's cognition. Therefore, children who master the classroom content (passage-level data) are expected to be able to read the extended books at the same level. In order to investigate whether the trained passage-level readability assessment models can be directly applied to books, we conduct experiments by training on passages and then perform transfer inference on books.

### 3.1 Datasets

**Passage-level Data** Following previous work on passage-level readability assessment, we collected texts in different grades from the textbooks of primary school (TPS) of more than ten publishers, where we deleted poetry and traditional Chinese texts. According to the specification in the *Chinese Curriculum Standards for Compulsory Education* by Ministry of Education, texts of Grade 1-2, Grade 3-4, Grade 5-6 are assigned with the difficulty level 1, 2, 3, respectively.

**Book-level Data** For book-level readability assessment, we constructed a dataset Graded Children's Books (GCB) in Chinese. We collected 320 books including popular stories and famous novels, with a total of 18 million characters. Each book is assigned with a suitable grade for children to read, according to the *Reading Guidance for School Students by Ministry of Education* [0] and *Basic Bibliography of the Library of Aiyue Primary School* [1]. If there is a conflict for a the same book, the standards of the Ministry of Education shall prevail. The difficulty is also classified into 3 levels, corresponding to Grade 1-2, 3-4, 5-6, respectively.

The statistics of both datasets is shown in Table 1. There are two notable differences between two datasets: 1) The number of book samples is much smaller than that of passage samples. 2) The average length of book texts is much longer than that of passage texts. These two properties pose more challenges for book readability assessment.

| Data | Level | Items | Avg.Len | Min.Len | Max.Len |
|---|---|---|---|---|---|
| TPS | 1 | 814 | 266 | 15 | 900 |
| | 2 | 1063 | 679 | 110 | 3903 |
| | 3 | 1104 | 1140 | 109 | 6054 |
| | All | 2981 | 737 | 15 | 6054 |
| GCB | 1 | 66 | 12935 | 175 | 154517 |
| | 2 | 100 | 48498 | 355 | 203041 |
| | 3 | 154 | 84391 | 339 | 317476 |
| | All | 320 | 58437 | 175 | 317476 |

Table 1: Statistics of the passage-level dataset TPS and the book-level dataset GCB.

| Model | Dataset | Precision | Recall | F1 |
|---|---|---|---|---|
| SVM | TPS | 87.231 | 87.333 | 87.140 |
| | TPS-GCB | 52.707 | 57.812 | 50.377 |
| TextCNN | TPS | 78.378 | 78.333 | 78.338 |
| | TPS-GCB | 53.469 | 49.375 | 38.965 |
| BERT | TPS | 83.852 | 83.667 | 83.236 |
| | TPS-GCB | 56.863 | 52.812 | 44.030 |

Table 2: Transfer inference results from passage-level TPS to book-level GCB.

### 3.2 Methods

We implement several classical classification models to predict the difficulty level of passages and books. We train models based on passage data and then test on both passage-level and book-level data. Specifically, on the passage data, the training, validation and testing set is split by 8:1:1. And for book-level transfer evaluation, all 320 books are considered as the test data.

**SVM**. The linguistic difficulty features are extracted using zhfeat (Li et al., 2022) which are normalized using MinMaxScaler (range -1 to 1). We adopt the libsvm [2] framework for our experiments.

[0] http://www.moe.gov.cn/jyb_xwfb/gzdt_gzdt/s5987/202004/W020200422556593462993.pdf
[1] http://www.cptoday.cn/news/detail/8309
[2] https://www.csie.ntu.edu.tw/~cjlin/libsvm/

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1298-1309, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

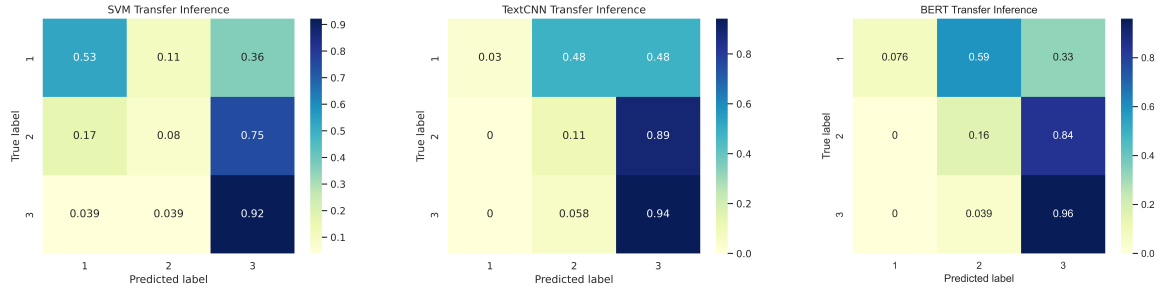



Figure 2: The confusion matrix of transfer inference from TPS to GCB, where the values on the diagonal are the recall of each category.

**TextCNN** (Kim, 2014). The maximum truncated length of the text is set to 2000. We adopt a random initialized character embedding. The number of convolution kernels is 128, and the size list is [2,3,4].

**BERT**. The maximum truncated length of the text is set to 2000 tokens, which is cut into 4 segments of 500. We use a sliding window to obtain the $[CLS]$ vector of each segment, and utilize the maximum pooling to obtain the final text representation.

### 3.3 Results

The experimental results are shown in Table 2. In terms of F1, SVM, TextCNN and BERT have a drop of 36.763 points (87.140-50.377), 39.373 points (78.338-38.965) and 39.206 points (83.236-44.030), respectively. These results demonstrate that the passage-level readability assessment models can't be directly applied to predict the difficulty level of a book. It is necessary to train a model specialized for book-level texts.

To display how the models perform predictions on different difficulty levels, we construct confusion matrices of transfer inference, as shown in Figure 2. Compared with SVM, TextCNN and BERT are sensitive to length and tend to predict a book as a higher difficulty level. We think the bias comes from the direct transfer inference. When we apply a model trained on passages that are relatively short to long texts, the model tends to predict the text as a higher difficulty level. As a result, the model obtains a quite high recall on the highest level but a low recall for the other levels.

## 4 Readability Assessment of Long Texts

To handle the book-level long text readability assessment, we propose a DSDR model and the overall structure is shown in Figure 3. Based on the Transformer structure, DSDR adds two modules: difficulty-aware segment pre-training and difficulty multi-view representation.

### 4.1 Difficulty-aware Segment Pre-training

Due to hardware limitations, we cannot directly utilize pre-trained models to process ultra long text. To make full use of the long text without loss of information, we compress a long text into multiple parts that are shorter and can be input into pre-trained models.

A book text is split into several segments, and the reading difficulty of each segment is different, which is also not consistent with the label of the whole book. To obtain a difficulty label for each segment, we classify the segments by clustering according to their linguistic features. Further, we leverage the automatic clustered data to do supervise training, through which the pre-trained model is enhanced with difficulty knowledge.

Concretely, given a book $B$, we split it into several fixed-length segments $B = (S_1, S_2, ..., S_N)$ to facilitate processing. We perform this segmentation on all books in the data, with $G$ denoting the total number of segments of all books.

Then for each segment $S_i$, we extract its linguistic difficulty features $F_i$:

$$F_i = \text{Extractor}(S_i), i = 1, 2, ..., G \quad (1)$$

CCL 2024

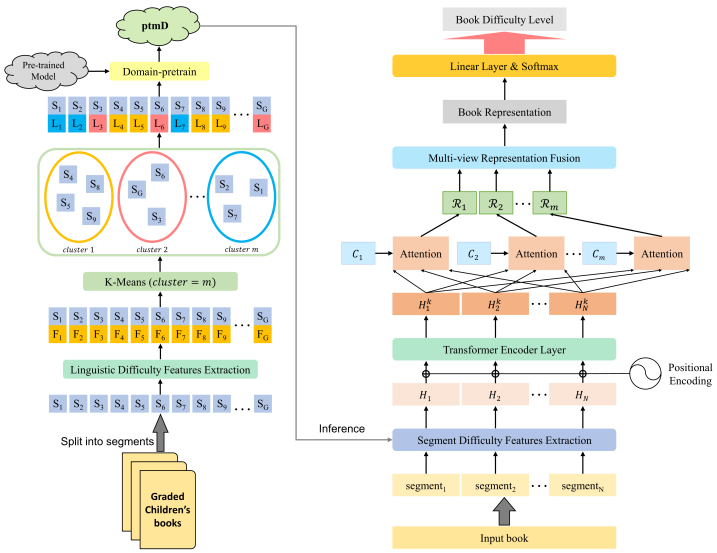



Figure 3: The overall architecture of our proposed DSDR model for book readability assessment.

where Extractor is the feature extractor. Following the previous work (Li et al., 2022), we adopt its Chinese linguistic features, including character, word and sentence features.

By extracting difficulty features in each segment of all the books, we get a series of pairs $(S_i, F_i)$ . Then we exploit the clustering algorithm KMeans to classify segments into $m$ categories:

$$L_i = \text{KMeans}(\text{clusters} = m, F_i), i = 1, ..., G \tag{2}$$

where $L_i$ is the category label of Segment $S_i$.

In this way, each segment is assigned with a unique label. Gathering all the segments, we construct a dataset $D = \{(S_1, L_1), (S_2, L_2), ..., (S_G, L_G)\}$, in an unsupervised way.

Further, on the dataset $D$, we employ the pre-trained model BERT followed by a linear layer to do a supervised training, where the training objective is to predict the difficulty label of each segment. Consequently, we obtain a **p**re-**t**rained **m**odel enhanced with **d**ifficulty knowledge:

$$\textbf{ptmD} = \text{BERT}(\text{train} = D) \tag{3}$$

### 4.2 Difficulty Multi-view Representation

Given an input book, we split it into $N$ segments $B = (S_1, S_2, ..., S_N)$, and then do inference on each segment with **ptmD** to get its text representation:

$$H_i = \textbf{ptmD}(S_i), i = 2, ..., N \tag{4}$$

where $H_i$ is the [CLS] token representation of the segment $S_i$.

Thus, for a book $B$, its semantic representation with difficulty knowledge is $H = (H_1, H_2, ..., H_N)$. In this way, a long text is compressed and represented with $H$ without loss of difficulty information.

Since the semantics of each segment is extracted independently, context information is lacking. Therefore, we apply the Transformer (Vaswani et al., 2017) encoder to supplement the contextual representation:

$$H^0 = H + P \tag{5}$$

$$H^t = \text{Transformer}(H^{t-1}), t = 1, 2, ..., k \tag{6}$$

where $P = (P_1, P_2, ..., P_N)$ is the positional encoding sequence, $P_i$ denoting the position of the $i$-$th$ segment in the book. $k$ is the number of encoder layers. The calculation method is consistent with the paper (Vaswani et al., 2017).

CCL 2024

Intuitively, a long text might contain information of various difficulty levels. To further exploit the difficulty knowledge, we set multiple difficulty representation vectors $C \in \mathbb{R}^{m \times d}$, where $d$ is the hidden size, and $m$ is the number of difficulty levels that is the same as the number of clusters set in KMeans.

Each difficulty representation vector $C_i \in \mathbb{R}^{1 \times d}$ is trained to extract the information of the corresponding difficulty level. The difficulty representation vectors $C$ are utilized to extract information from $H^k$ using cross-attention. The extracted information $R \in \mathbb{R}^{m \times d}$ are representations of $m$ difficulty levels:

$$R = \text{Attention}(C\text{W}^{\text{Q}}, H^k\text{W}^{\text{K}}, H^k\text{W}^{\text{V}}) \tag{7}$$

$$\text{Attention}(\text{Q}, \text{K}, \text{V}) = \text{Softmax}(\frac{\text{QK}^{\text{T}}}{\sqrt{\text{d}}})\text{V} \tag{8}$$

where $\text{W}^{\text{Q}}$, $\text{W}^{\text{K}}$ and $\text{W}^{\text{V}}$ are trainable parameters. $R = (R_1, R_2, R_3, ..., R_m) \in \mathbb{R}^{m \times d}$ is the difficulty multi-view representation, where $R_i$ is the singe-view representation of segment difficult Level $i$, which is not directly mapped to the book's difficulty level.

### 4.3 Difficulty Level Prediction

Based on multi-view representations, we employ mean pooling to obtain the overall representation:

$$T = \text{MeanPooling}(R_1, R_2, R_3, ..., R_m) \tag{9}$$

Finally, we compute the probability that a book belongs to the $i$-$th$ difficulty category by:

$$p_i = \text{Softmax}(T\text{W} + \text{b}) \tag{10}$$

where W are trainable parameters, and b are scalar biases.

## 5 Experimental Setup

### 5.1 Data and Metrics

We conduct experiments on our new data GCB, consisting of 320 children's books labelled with difficulty levels. The statistics of GCB is listed in Table 1. We apply stratified k-fold (k=5, train=0.8, test=0.2) and report the average results.

For evaluation, we calculate the weighted precision, recall, F1 score and accuracy.

### 5.2 Comparing Methods

We provide a comprehensive comparison with previous approaches, including the traditional SVM, the neural network classifiers (TextCNN, DPCNN) and the pre-trained models (BERT, BigBird).

**SVM-L**. Only the length features are input into the SVM classifier.

**SVM**. **TextCNN**. **BERT**. The details are the same as in Section 3.2, except for the maximum length in TextCNN and BERT.

**DPCNN** (Johnson and Zhang, 2017). It increases the number of layers of CNN to extract long-range text dependencies. The number of convolution kernels is 128.

**BERT-FP-LBL** (Li et al., 2022). We apply the most recent passage-level model on Chinese data, where the adjustment factor $\rho$ is 0 and other parameters remain the same as in the original paper.

**BigBird** (Zaheer et al., 2020). It uses the sparse attention to handle sequences that are eight times longer than usual.

**PyramidFormer**. Inspired by the work (Cirstea et al., 2022), we designed a model for long sequence compression, in which we replace the self-attention in BERT with the linear patch attention. We use random initialized character embeddings.

**DS+TextCNN**. **DS+DPCNN**. We apply our **ptmD** model to get the difficulty-aware text representation $H$, and use $H$ as the input sequence of TextCNN and DPCNN, respectively. Other settings remain unchanged.

CCL 2024

### 5.3 Training Details

We do experiments using the Pytorch (Paszke et al., 2019) framework. For training, we use the AdamW optimizer, the weight decay is 0.02 and the warm-up ratio is 0.1. For the difficulty-aware segment pre-training in Section 4.1, the number of difficulty categories $m$ is 14, the segment length is 500, the batch size is 16, the epoch is 10 and the learning rate is 3e-5. For the difficulty multi-view representation in Section 4.2, the batch size is 16, the number of Transformer encoder layers k is 3, the epoch is 100 and the learning rate is 3e-5. We choose the length as 180000 which covers 95% of the books. Accordingly, the book is represented as 360 (180000/500) segments, as shown in Table 3.

## 6 Results and Analysis

| Model | Max.Len | Embedding | Precision | Recall | F1 | Accuracy |
|---|---|---|---|---|---|---|
| SVM-L | Unlimited | - | 62.908 | 66.250 | 63.002 | 66.250 |
| SVM | Unlimited | - | 68.587 | 68.750 | 68.228 | 68.750 |
| TextCNN | 180000 | 180000 × 64 | 58.488 | 55.937 | 46.401 | 55.937 |
| DPCNN | 180000 | 180000 × 64 | 68.860 | 68.750 | 68.142 | 68.750 |
| BERT | 8192 | 16 × 512 × 768 | 50.258 | 61.875 | 54.309 | 61.875 |
| BERT-FP-LBL | 4000 | 8 × 500 × 768 | 61.420 | 63.437 | 57.178 | 63.437 |
| BigBird | 4096 | 4096 × 768 | 53.710 | 61.250 | 51.146 | 61.250 |
| PyramidFormer | 180000 | 180000 × 64 | 69.061 | 69.062 | 68.253 | 69.062 |
| DS + TextCNN | 180000 | 360 × 768 | 69.306 | 69.062 | 67.997 | 69.062 |
| DS + DPCNN | 180000 | 360 × 768 | 72.338 | 70.625 | 70.831 | 70.625 |
| DSDR | 180000 | 360 × 768 | **73.266** | **73.125** | **72.466** | **73.125** |

Table 3: Overall experimental results on children's books for readability assessment.

### 6.1 Overall Results

The overall experimental results are summarized in Table 3. Our proposed DSDR model achieves the best performance on all metrics. Compared with the traditional SVM classifier, our model improves by 4.238 on F1. The simple SVM classifier with only the length feature (SVM-L) obtains an F1 score of 63.002, which is much better than TextCNN, BERT and BigBird, demonstrating that the length feature plays a vital role for readability assessment. But it is about 10 points lower than our model DSDR.

Though the pre-trained models are becoming mainstream for various NLP tasks, it is not feasible to directly apply them to our task. Using the pretrained model BERT only obtains an F1 score of 54.309. Even for the pre-trained model specialized for long text like BigBird, it's too long of our book text to be compressed as an input, resulting in a poor result of 51.146 F1 score. In contrast, our model DSDR processes the long input sequence effectively and at the same time incorporates difficulty knowledge, outperforming BigBird by more than 20 points on F1 and more than 10 points on accuracy.

Compared with TextCNN and DPCNN, both DS+TextCNN and DS+DPCNN bring big improvements by applying our trained **ptmD** to get the difficulty-aware representation as model input. It improves F1 by 21.596 points over TextCNN, and outperforms the advanced DPCNN by 2.689 F1 points. This validates that our difficulty-aware segment pre-training is flexible, and can be easily applied to other neural network models to improve readability assessment performance.

### 6.2 Ablation Study

We explore the contribution of each component in our DSDR model by conducting ablation studies with the following settings:

**-DPT**: Do not apply difficultty-aware segment pre-training, instead, we use the original pretrained model to extract difficulty features from segments.

**-MVDR**: Do not use multi-view representations.

The results are shown in Table 4. Without applying difficulty-aware segment pre-training leads to an obvious drop of performance, with about 5 points on both F1 and accuracy. In contrast, after exploiting difficulty-aware segment pre-training on our automatically constructed data, the pre-trained model learns to extract segment difficulty features.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1298-1309, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

| Model | F1 | Acc |
|---|---|---|
| DSDR | 72.466 | 73.125 |
| -DPT | 67.402 | 68.125 |
| -MVDR | 71.142 | 71.563 |

Table 4: Ablation study on the model structure.

Without the use of difficulty multi-view representations (MVDR) also leads to a drop of performance. This indicates that, for a book-level long text with varying difficulty levels, it's important to set up multiple difficulty representations.

### 6.3 Confusion Matrix

To clearly display the difficulty levels predicted by different models, we construct confusion matrices based on the predicted results of SVM, BERT, BigBird and DSDR. The results are shown in Figure 4. Compared with SVM, our DSDR model obtains an obvious performance gain on Level 1, and compared with BERT and BigBird, our model achieves significant improvements on Level 2.

As shown in Figure 4, BERT and BigBird obtain low recall value on Level 2 and high recall value on Level 3. For BERT and Bigbird, we need to truncate the input text into a certain length to feed into the pre-trained model, which loses length information.

After truncation, the texts of Level 2 and Level 3 have similar lengths. As a result, the model is likely to predict texts of Level 2 as Level 3, which causes the low recall in Level 2 and the false high recall in Level 3. In contrast, our model represents the long text with multiple difficulty-enriched segments, avoiding the loss of information.

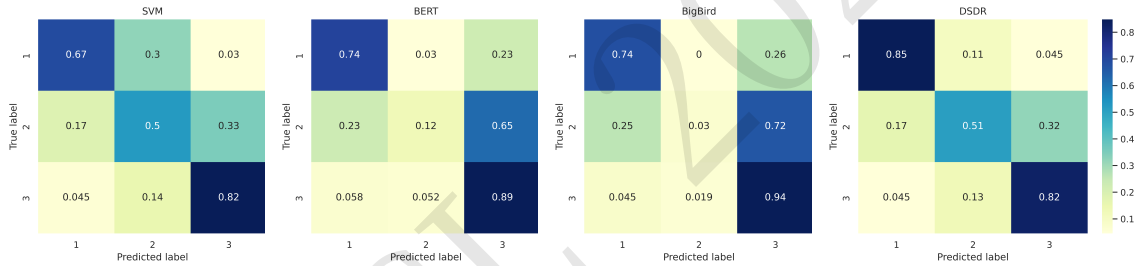



Figure 4: Comparative analysis of confusion matrix.

### 6.4 Analysis on the Number of Clusters

During difficulty-aware segment pre-training, we exploit the clustering algorithm KMeans to classify different segments into multiple categories. To explore the effect of the number of clusters on the results, we set up different $m$ and report the experimental results in Figure 5.

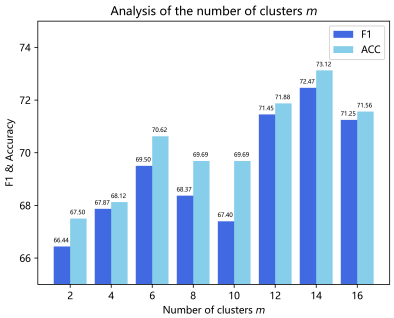



Figure 5: Effect of the the number of clusters $m$.

When $m$ is small, the model's ability to distinguish difficulty levels is weak. When $m$ is relatively big, more difficulty fragments are extracted, and the overall performance of modeling as multiple representations is better. When $m$=14, we obtain the best results on both F1 and accuracy.

### 6.5 Experiments on Different Multi-view Representation Fusion Methods

With reference to the work (Kong et al., 2022), we design several different ways of difficulty multi-view representation fusion.

**Average Weight (AW)**. The method used in our paper.

**Random Weight (RW)**. We set a learnable parameter $\alpha \in \mathbb{R}^{1\times m}$, and compute the weight of each difficulty representation:

$$T = \sum_{i=1}^{m} \frac{e^{\alpha_i}}{e^{\alpha_1} + \cdots + e^{\alpha_m}} R_i \tag{11}$$

**Transforming Weight (TW)**. We set a linear layer (LL) to transform each difficulty representation to 1 dimension, and compute the weights by:

$$T = \sum_{i=1}^{m} \frac{\mathrm{LL}(R_i)}{\mathrm{LL}(R_1) + \cdots + \mathrm{LL}(R_m)} R_i \tag{12}$$

The experimental results are shown in Table 5. The fusion method does not have a significant effect on the results. We believe this is because our difficulty vector $C$ is learned as a parameter and is able to adapt itself to changes in the weights of individual difficulty representations, and thus it is not sensitive to fusion methods.

| Method | P | R | F1 | Acc |
|---|---|---|---|---|
| AW | 73.266 | 73.125 | 72.466 | 73.125 |
| RW | 72.738 | 73.125 | 72.147 | 73.125 |
| TW | 73.464 | 72.500 | 72.541 | 72.500 |

Table 5: Experiments on different fusion methods.

### 6.6 Case Study

To display our interesting task of readability assessment for children's books, we list some examples in Table 6, reporting the predicted results by DSDR and some baseline models.

| Title | Length | True level | Predict Label | | | |
|---|---|---|---|---|---|---|
| | | | SVM | BERT | BigBird | DSDR |
| 绿山是一个谜(Green Mountain is a mystery) | 11191 | 1 | 1 | 3 | 3 | 1 |
| 小飞人卡尔松 (Peter Pan Karlsson) | 154340 | 2 | 3 | 3 | 3 | 2 |
| 帽子的秘密 (The Secret of the Hat) | 50805 | 2 | 3 | 3 | 3 | 2 |
| 小兵张嘎 (Little Soldier Zhang Ga) | 54901 | 3 | 2 | 3 | 3 | 3 |
| 布鲁克林有棵树 (A Tree Grows in Brooklyn) | 277362 | 3 | 1 | 3 | 3 | 2 |

Table 6: Case study

## 7 Conclusion

In this paper, we propose a novel model DSDR to assess the reading difficulty of book-level long texts. Through difficulty-aware segment pre-training, the pre-trained model is enhanced with the ability to represent difficulty knowledge, and the difficulty multi-view representations help to compress a long text into a representation without loss of information. In addition, we construct a new dataset Graded Children's Books in which each book is labelled with a difficulty level. Experimental results show that our approach achieves superior performance against traditional methods and popular pre-trained language models.

CCL 2024

## Acknowledgements

This work is supported by the Key Project of Natural Science Foundation of China (61936012) and the National Natural Science Foundation of China (62076008).

## References

Ion Madrazo Azpiazu and Maria Soledad Pera. 2019. Multiattentive recurrent neural network architecture for multilingual readability assessment. *Transactions of the Association for Computational Linguistics*, 7:421–436.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. 2020. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*.

Razvan-Gabriel Cirstea, Chenjuan Guo, Bin Yang, Tung Kieu, Xuanyi Dong, and Shirui Pan. 2022. Triformer: Triangular, variable-specific attentions for long sequence multivariate time series forecasting–full version. *arXiv preprint arXiv:2204.13767*.

Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

Tovly Deutsch, Masoud Jasbi, and Stuart Shieber. 2020. Linguistic features for readability assessment. *arXiv preprint arXiv:2006.00377*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2019. Poly-encoders: Transformer architectures and pre-training strategies for fast and accurate multi-sentence scoring. *arXiv preprint arXiv:1905.01969*.

Joseph Marvin Imperial. 2021. Bert embeddings for automatic readability assessment. *arXiv preprint arXiv:2106.07935*.

Zhiwei Jiang, Qing Gu, Yafeng Yin, and Daoxu Chen. 2018. Enriching word embeddings with domain knowledge for readability assessment. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 366–378.

Rie Johnson and Tong Zhang. 2017. Deep pyramid convolutional neural networks for text categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 562–570.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Conference on Empirical Methods in Natural Language Processing*.

Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*.

Weize Kong, Swaraj Khadanga, Cheng Li, Shaleen Kumar Gupta, Mingyang Zhang, Wensong Xu, and Michael Bendersky. 2022. Multi-aspect dense retrieval. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3178–3186.

Justin Lee and Sowmya Vajjala. 2022. A neural pairwise ranking model for readability assessment. *arXiv preprint arXiv:2203.07450*.

Bruce W Lee, Yoo Sung Jang, and Jason Hyung-Jong Lee. 2021. Pushing on text readability assessment: A transformer meets handcrafted linguistic features. *arXiv preprint arXiv:2109.12258*.

James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontanon. 2021. Fnet: Mixing tokens with fourier transforms. *arXiv preprint arXiv:2105.03824*.

CCL 2024

Wenbiao Li, Ziyang Wang, and Yunfang Wu. 2022. A unified neural network model for readability assessment with feature projection and length-balanced loss. *arXiv preprint arXiv:2210.10305*.

Shizhan Liu, Hang Yu, Cong Liao, Jianguo Li, Weiyao Lin, Alex X Liu, and Schahram Dustdar. 2021. Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting. In *International Conference on Learning Representations*.

Dawei Lu, Xinying Qiu, and Yi Cai. 2019. Sentence-level readability assessment for l2 chinese learning. In *Workshop on Chinese Lexical Semantics*, pages 381–392. Springer.

Yi Ma, Eric Fosler-Lussier, and Robert Lofthus. 2012. Ranking-based readability assessment for early primary children's literature. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 548–552.

G Harry Mc Laughlin. 1969. Smog grading-a new readability formula. *Journal of reading*, 12(8):639–646.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Maria Soledad Pera and Yiu-Kai Ng. 2014. Automating readers' advisory to make book recommendations for k-12 readers. In *Proceedings of the 8th ACM Conference on Recommender Systems*, pages 9–16.

Subha Perni, Michael K Rooney, David P Horowitz, Daniel W Golden, Anne R McCall, Andrew J Einstein, and Reshma Jagsi. 2019. Assessment of use, specificity, and readability of written clinical informed consent forms for patients with cancer undergoing radiotherapy. *JAMA oncology*, 5(8):e190260–e190260.

Sarah E Petersen and Mari Ostendorf. 2009. A machine learning approach to reading level assessment. *Computer speech & language*, 23(1):89–106.

Xinying Qiu, Kebin Deng, Likun Qiu, and Xin Wang. 2017. Exploring the impact of linguistic features for chinese readability assessment. In *National CCF Conference on Natural Language Processing and Chinese Computing*, pages 771–783. Springer.

Xinying Qiu, Yuan Chen, Hanwu Chen, Jian-Yun Nie, Yuming Shen, and Dawei Lu. 2021. Learning syntactic dense embedding with correlation graph for automatic readability assessment. *arXiv preprint arXiv:2107.04268*.

Lloyd Lois Antonie Reyes, Michael Antonio Ibañez, Ranz Sapinit, Mohammed Hussien, and Joseph Marvin Imperial. 2022. A baseline readability model for cebuano. *arXiv preprint arXiv:2203.17225*.

Antony Sare, Aesha Patel, Pankti Kothari, Abhishek Kumar, Nitin Patel, and Pratik A Shukla. 2020. Readability assessment of internet-based patient education materials related to treatment options for benign prostatic hyperplasia. *Academic Radiology*, 27(11):1549–1554.

Sarah E Schwarm and Mari Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd annual meeting of the Association for Computational Linguistics (ACL'05)*, pages 523–530.

Kathleen M Sheehan, Irene Kostin, Yoko Futagi, and Michael Flor. 2010. Generating automated text complexity classifications that are aligned with targeted text complexity standards. *ETS Research Report Series*, 2010(2):i–44.

Chao-Hong Tan, Qian Chen, Wen Wang, Qinglin Zhang, Siqi Zheng, and Zhen-Hua Ling. 2021. Ponet: Pooling network for efficient token mixing in long sequences. *arXiv preprint arXiv:2110.02442*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. 2020. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*.

Zhenhui Xu, Meng Zhao, Liqun Liu, Lei Xiao, Xiaopeng Zhang, and Bifeng Zhang. 2022. Mixture of virtual-kernel experts for multi-objective user profile modeling. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4257–4267.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33:17283–17297.

CCL 2024

# Mitigating the Bias of Large Language Model Evaluation

**Hongli Zhou**[1], **Hui Huang**[2], **Yunfei Long**[3], **Bing Xu**[2], **Conghui Zhu**[2],
**Hailong Cao**[2], **Muyun Yang**[2*], **Tiejun Zhao**[2]
[1]School of Architecture and Design, Harbin Institute of Technology, Harbin, China
[2]Faculty of Computing, Harbin Institute of Technology, Harbin, China
[3]University of Essex
{hongli.joe,huanghui}@stu.hit.edu.cn;yl20051@essex.ac.uk;
{hitxb,conghui,caohailong,yangmuyun,tjzhao}@hit.edu.cn

## Abstract

Recently, there has been a trend of evaluating the Large Language Model (LLM) quality in the flavor of LLM-as-a-Judge, namely leveraging another LLM to evaluate the current output quality. However, existing judges are proven to be biased, namely they would favor answers which present better superficial quality (such as verbosity, fluency) while ignoring the instruction following ability. In this work, we propose systematic research about the bias of LLM-as-a-Judge. Specifically, for closed-source judge models, we apply calibration to mitigate the significance of superficial quality, both on probability level and prompt level. For open-source judge models, we propose to mitigate the bias by contrastive training, with curated negative samples that deviate from instruction but present better superficial quality. We apply our methods on the bias evaluation benchmark, and experiment results show our methods mitigate the bias by a large margin while maintaining a satisfactory evaluation accuracy.

## 1 Introduction

Recently, the evaluation for Large-scale Language Models (LLMs) has drawn considerable attention of research community (Liang et al., 2022; Chang et al., 2023). As the capabilities of LLMs continue to develop across various tasks, it is essential to evaluate them from a comprehensive perspective (Qin et al., 2023). However, traditional evaluation metrics for generative models, such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004), only capture limited aspects of a model's performance.

Some research has proposed LLM-as-a-Judge (Li et al., 2023; Zheng et al., 2023), namely utilizing proprietary LLMs, especially GPT-4 (Achiam et al., 2023), to evaluate the LLM's response. However, relying on external API for evaluation may introduce consideration about privacy leakage. Other works propose to fine-tune open-source models as external evaluators (Wang et al., 2024; Zhu et al., 2023). They construct evaluation data and fine-tune LLMs to endow them specialized evaluation ability comparable with GPT-4. However, the fine-tuned judges underperform GPT-4 in terms of fairness and scalability by a large margin (Huang et al., 2024).

As a pioneering research, LLMBar (Zeng et al., 2023) introduces the first meta-evaluation benchmark of LLM evaluation bias. The benchmark is adversarially designed to test whether LLM evaluators can detect instruction-following outputs. They apply their benchmark on existing LLM evaluators, and reveal that even state-of-the-art LLM evaluators struggle to provide unbiased evaluation on their benchmark. However, their study is mainly based on closed-source models, and they only provide several suggestions to mitigate the bias by prompt engineering.

In this work, we propose a systematic framework to mitigate the bias of LLM evaluators. For closed-source LLM judges, we propose to mitigate the bias by calibration. As the evaluation bias is generally presented as a higher score for superficial quality, we propose two ways to directly model the superficial quality, and subtract it from the final result. More specifically, for probability-based evaluators, we propose to model the superficial quality by utilizing the probability generated by the pre-trained model.

*Corresponding author.
©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1310-1319, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers

For generation-based evaluators, we propose to model the superficial quality by probing the superficial attributes in the prompt template. For open-source LLM evaluators, we propose to mitigate the bias of LLM evaluators by pairwise contrastive training. Delicately curated negative samples are included in the training set where the answer deviates from the instruction but presents better superficial quality.

We evaluate our methods on the bias benchmark of LLMBar, which consists of one natural and four adversarial test sets to benchmark LLM evaluators' ability to detect instruction-following answers unbiasedly. Experiment results show that all of our proposed methods mitigate the evaluation bias on the adversarial sets by a large margin, while keeping a comparable evaluation performance on the natural set.

## 2 Related Work

Presently, evaluation research on LLMs can be grouped into three distinct methodologies: human evaluation, benchmark evaluation, and LLM-as-a-Judge.

The most straightforward approach to evaluating an LLM involves engaging dedicated evaluators or crowdsourcing to compare and assess the model's output (Guo et al., 2023; Bang et al., 2023; Bubeck et al., 2023). However, human evaluation is labor-intensive and challenging to scale during the model development cycle. Additionally, human subjective factors can influence the evaluation results, making it difficult to guarantee the consistency and stability of model evaluations. Other scholars have devised standard test sets for the automatic evaluation of LLMs. These test sets typically encompass a range of tasks, assessing the LLM's performance by comparing its task outputs to standard answers (Liang et al., 2022; Yu et al., 2024; Zhong et al., 2023). However, the benchmarks typically consist of problems of selection, which are not adapted to quantify the generation ability of LLMs.

Given the various issues with benchmark evaluation, some scholars have introduced the evaluation scheme with the name of LLM-as-a-Judge, which involves utilizing another proprietary LLMs to assess the output of the current model. For example, AlpacaEval (Li et al., 2023) constructs an evaluation set comprising 805 question-answer pairs and leverages GPT-4 to compare the winning rate of the model's answers against text-davinci-003. Similarly, MT-bench (Zheng et al., 2023) creates 80 high-quality multi-round test questions spanning eight common areas, including mathematics and reasoning. It then automatically grades the model's answers using GPT-4.

However, relying on an external model to perform evaluation may introduce concerns about the privacy leakage. Also, the iterative versions of the API model also make the evaluation inconsistent. To tackle these challenges, some studies have advocated for the training of localized agent evaluation models. For instance, PandaLM (Wang et al., 2024) constructs data based on Alpaca instructions and GPT-3.5 annotation, and then fine-tunes LLaMA-7B (Wang et al., 2024) as a judge model. JudgeLM (Zhu et al., 2023) constructs data from GPT-4 annotations and fine-tunes a scalable judge model.

Despite the LLM-as-a-Judge becoming the default paradigm for LLM evaluation, various work has pointed out that the LLM-based judges are severely biased, namely the evaluator would make prediction based on spurious features (mainly the superficial quality) while ignoring the alignment with the instruction. Various biases are discussed in previous work, including position bias (Wang et al., 2023), verbosity bias (Saito et al., 2023), formality bias and self-enhancement bias (Zheng et al., 2023), etc. As a pioneering work, LLMBar proposes the first systematic evaluation of LLM evaluation bias. They created four adversarial test sets as the testbed for LLM's ability of making unbiased prediction. Their results indicate that LLM evaluators, even GPT-4, suffer severely from the over-reliance on the superficial quality. However, they only propose to compensate for the bias by prompt engineering, which can hardly be applied to fine-tuned judges. A more thorough bias mitigation framework is of urgent demand.

## 3 Methodology

### 3.1 The Formal Definition of LLM-as-a-Judge

Given a pre-defined evaluation set which contains various instructions, the LLM can be evaluated in both pairwise and pointwise manners:

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1310-1319, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

$$\mathrm{P}(A|I) = argmax \sum_{t=m}^{m+n} \log p(h_t|\mathrm{T}(I, A), \theta)$$

$$\mathrm{P}(A_1, A_2|I) = argmax \sum_{t=m}^{m+n} \log p(h_t|\mathrm{T}(I, A_1, A_2), \theta)$$

where $h_m$ to $h_{m+n}$ is the span of predicted tokens, $T(\cdot)$ denotes the prompt template which is designed according to the judge and the evaluation scheme, $I$ denotes the instruction, $A$ denotes the answer, and $\theta$ is the model parameters of LLM. When generating an evaluation through decoding, the token of the evaluation part will be generated by taking the argmax of the vocabulary probability of each step.

Moreover, denoted by (Fu et al., 2023), the generated log-probabilities (if available), is also an informative quality indicator that can be leveraged for evaluation:

$$\mathrm{P}(A|I) = \sum_{t=m}^{m+n} w_t \log p(h_t|h_{<t}, \mathrm{T}(I, A), \theta)$$

where $h_m$ to $h_{m+n}$ is the span of target tokens, and $w_t$ is the weight of the token at position $t$.

As the LLM-as-a-Judge can rely on both general-purposed closedq-source LLMs and specially fine-tuned closed-source LLMs, we propose two different strategies dedicated to mitigate the LLM evaluation bias in different scenarios. In the next sections, we will introduce the two methods one by one.

### 3.2 Online Mitigation by Calibration

As the bias of LLM evaluation mainly comes from the over-reliance to the superficial quality, the bias mitigation can be achieved by calibration in an on-the-fly manner. Calibration is a widely adopted technique to curate the language model predictions (Zhao et al., 2021). Neural autoregressive language models take a sequence of tokens as input and output a probability distribution over the next token. As LLMs are typically trained with maximizing the log-likelihood, its prediction tends to be biased towards tokens that are frequent in the training set. Therefore, the bias can be mitigated by firstly quantifying the bias lying in the prediction distribution, then make a explicit subtraction. More specifically, as discussed in Section 3.1, there are two flavors of LLM-as-a-Judge, namely probability-based and generation-based, and we propose different quantification methods for the superficial quality, respectively.

For probability-based evaluation, we propose to quantify the superficial quality with pre-trained models. The reason behind it is simple, since LLM typically gains its instruction following ability during supervised fine-tuning (SFT), the difference between the prediction distribution of pre-trained and SFT model can be indicative of instruction alignment. If one instruction-answer pair is favored by the pre-trained model, but unfavored by the instruction-tuned model, we would deem the answer as holding fluency but no instruction alignment.

For generation-based evaluation, as the prediction logits is inaccessible, it is hard to calculate the distribution difference between pre-trained and SFT models. Therefore, we propose to directly model superficial quality by prompt engineering. More specifically, we design three prompts for quantifying the superficial quality, as shown in Figure 1.

Finally, the attribute of superficial quality is excluded from the evaluation result by subtraction:

$$\hat{\mathrm{P}}(\mathrm{A}|\mathrm{I}) = Normalize(\mathrm{P}(A|I)) - \alpha Normalize(\mathrm{Q}(A|I)) \quad (1)$$

$$Normalize(x_i) = \frac{Max(x_i) - x_i}{Max(x_i) - Min(x_i)}$$

where $Q(A|I)$ denotes the scores derived from superficial quality modeling, and $\alpha$ is a balancing factor, and $Normalize(\cdot)$ is to normalize predictions of different models into the same scale. Intuitively, if an instruction-answer pair holds high superficial quality (such as fluency, verbosity, engaging tones, etc), but misaligns with the instruction, it may receive a higher score in $P(A|I)$, but also a higher score in $Q(A|I)$, therefore our calibrated evaluation would be less biased, as illustrated in Figure 2.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1310-1319, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

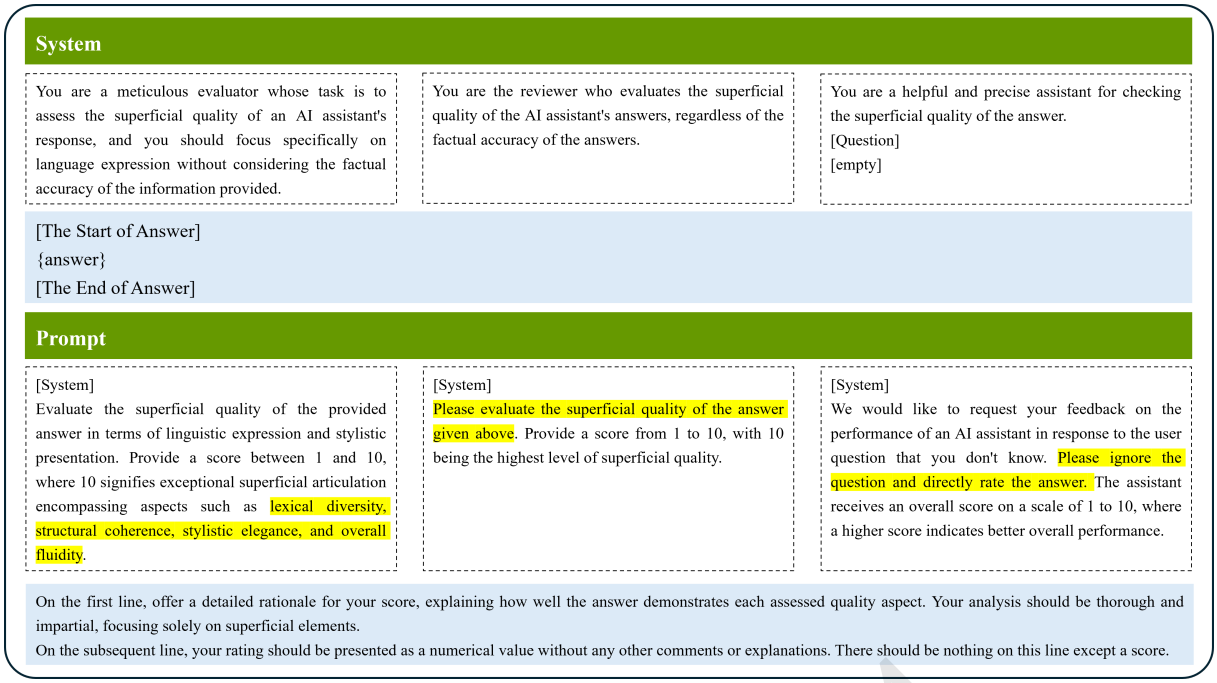



Figure 1: The three prompts for modeling superficial quality. The text in the blue blocks is the common part of the prompts, and together with the text in the dotted blocks, they constitute three types of prompts.

### 3.3 Offline Mitigation by Contrastive Training

For fine-tuned LLM judges, both the checkpoint and the training data are openly-available. Therefore, it is possible to mitigate the bias in an offline manner, by preventing the model from over-relying on the spurious features for prediction. More specifically, we apply contrastive training on the LLM judge with delicately curated adversarial negative samples. Our method consists of the following two steps, namely negative sample construction and contrastive training, as shown in Figure 3.

Firstly, we would like to create negative samples which are misaligned with the instruction, but hold better superficial quality. Given an instruction $I \in D$ where $D$ is its corresponding dataset, we retrieve a closely related yet sufficiently different instruction $I'$ from the same dataset D:

$$I' = \text{argmax}_{I' \in D, sim(I,I') < \epsilon} sim(I, I')$$

where $sim(\cdot, \cdot)$ is the cosine similarity measured by Instructor (Su et al., 2023), a sentence embedding model specifically designed for modeling instruction intentions, and $\epsilon$ is a threshold to ensure that $I'$ and $I$ are semantically different enough. Then, in combination with the preferred answer $A$, this process enables us to create triplets $(I, A, A')$, with the expectation that $A$ faithfully follows the instruction $I$, while $A'$ deviates from $I$.

Secondly, the judge model is trained based on the combination of both contrastive samples and natural samples. We followed the instruction tuning paradigm, format all datasets to follow a chatbot-style schema to unify the varied styles and formats of the instruction datasets. During this process, the judge model can both see natural triplets $(I, A, A')$ with the $A$ holds better superficial quality and better instruction following, and also see adversarial triplets $(I, A, A')$ with the $A'$ is more fluent, verbose but deviates from the instruction intentions, therefore the judge model would learn an unbiased prediction pattern, avoiding making prediction solely based on superficial quality. Notice during this process we does not introduce extra annotation, only constructing negative samples based on the available data, therefore our method is both effective and practicable.

CCL 2024

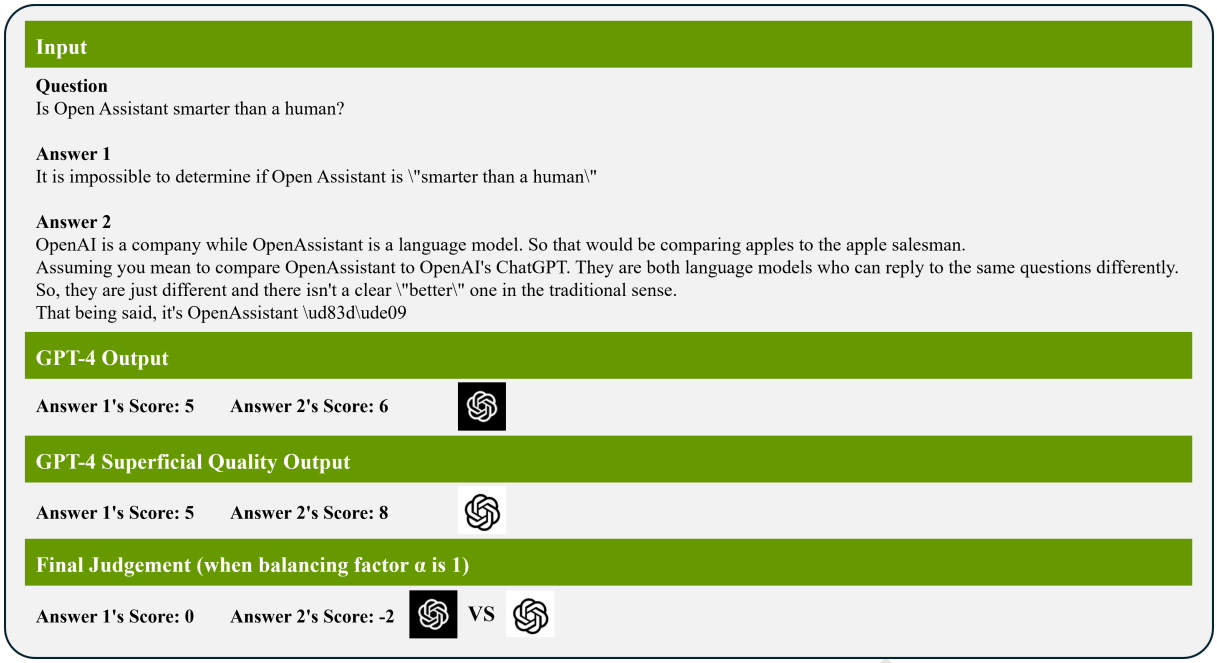



Figure 2: An example of online mitigation by calibration for generation-based evaluation. The correct output is answer 1 is better. It can be seen that the bias of LLM evaluation is effectively mitigated by subtracting the superficial quality.

## 4 Experiment

### 4.1 Set-up

Our experiments are based on LLMBar (Zeng et al., 2023), which consists a Natural set and four Adversarial sets. The Natural set collects and filters preference data from existing benchmarks, aiming to gauge evaluator performance in real-world distributions. Conversely, the Adversarial set comprises adversarially crafted instances that tend to confound less adept evaluators. There are four adversarial sets, namely Neighbor, GPTInst, GPTOut, Manual, which are created in different methods to probe the evaluator's bias from different perspectives.

For fine-tuning the open-source judges, we leverage the data from JudgeLM (Zhu et al., 2023), which contains 100K judge samples, with the instructions sampled from a large-scale set that contains various instruction datasets, and the answers are generated from leading open-source LLMs, and the evaluations are annotated by GPT-4. Due to the limitation of computational resource, we sample 20K samples for negative sample construction.

For online mitigation by calibration, we conduct experiments on two widely used models, namely Text-davinci-003[0] and GPT-4-1106-preview[1], both have gone through supervised fine-tuning and gained instruction following ability. For probability calibration, we use Davici-002 to quantify the superficial quality, which is the non-instruction tuned version of Text-davinci-003.

For offline mitigation by contrastive training, we follow the setting of JudgeLM (Zhu et al., 2023), which is a specifically fine-tuned judge model based on Vicuna-7B (Chiang et al., 2023). We conduct negative sample generation inside the dataset, therefore no extra annotation is required.

### 4.2 Main Results

As can be seen in Table 1, both of our proposed methods managed to achieve a higher performance on the adversarial sets, while keeping a decent evaluation performance on the natural set. This verifies that we succeed to mitigate the bias of LLM evaluation.

[0]https://platform.openai.com/docs/models/gpt-3-5
[1]https://platform.openai.com/docs/models/gpt-4-1106-preview

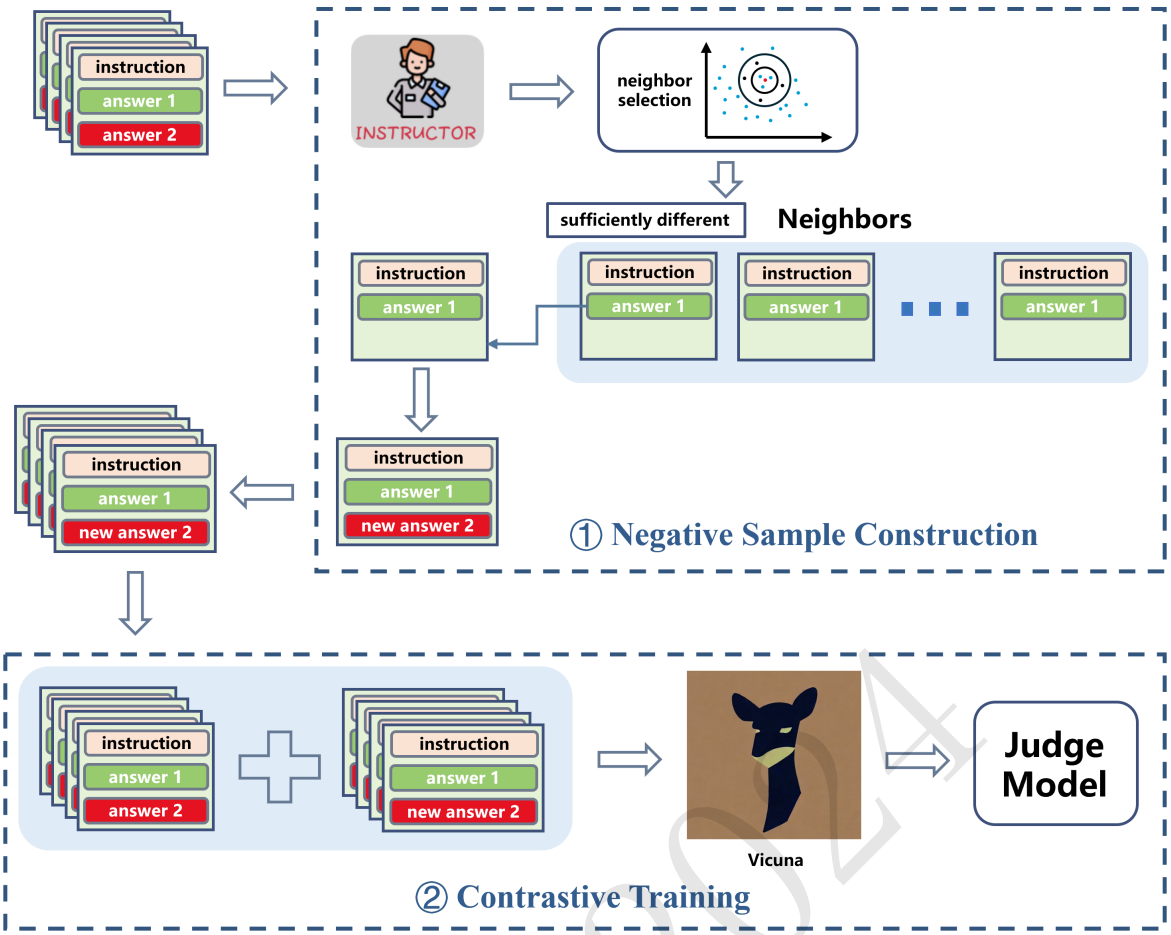



Figure 3: The pipeline of our offline mitigation. The first step is to construct negative samples in the original dataset D, and the second step is to perform contrastive training with the original data and the newly constructed data.

Between the two calibration methods, probability calibration achieves better performance, with consistent improvement on all of the four adversarial test sets, while leading no degradation to the original evaluation performance. Prompt calibration also achieves improvement on all adversarial test sets, but the improvement is comparably marginal, and the performance on the natural set is also slightly degraded. This indicates that the bias towards superficial quality mainly comes from the pre-training process, and can not be effectively detected solely based on the SFT model. However, it is worth noting that the accuracy of GPT-4 on the GPTOut set is largely improved, which demonstrates that the self-enhancement bias is mitigated. On the other hand, by calculating the probability distribution difference between the pre-trained and SFT models, we can precisely model the instruction alignment, thereby achieving superior performance on both natural and adversarial test sets.

For the contrastive training based results, we also achieve improvement on the adversarial test sets. The only exception is GPTOut, where the contrastive trained model achieves inferior performance. We think this might be due to the results constructed by GPTOut are generated by GPT-4, which presents a clear pattern that cannot be imitated by automatic negative sampling. On other words, this subset contains more formality bias, which deserves our future exploration.

It can also be notified that it is difficult to improve the evaluation accuracy based on bias mitigation, as the superficial quality is also an important classification feature. Answer with better detailedness, fluency and formality is also favored by the human preference, therefore excluding these aspects in evaluating LLM performance is actually unreasonable.

| Model | Method | Mitigation | Natural | Neighbor | Manual | GPTOut | GPTInst | Average |
|---|---|---|---|---|---|---|---|---|
| T003 | logprobs | N/A | 70.00 | 43.33 | 43.47 | 48.94 | 38.67 | 48.88 |
| | | prob-calib | **71.50** | **72.83** | **65.21** | **57.20** | **68.38** | **67.02** |
| T003 | genlogits | N/A | **76.00** | 31.68 | 28.57 | **51.35** | 27.33 | 42.99 |
| | | prob-calib | 72.50 | **69.27** | **72.27** | 48.94 | **71.35** | **66.87** |
| D002 | logprobs | N/A | 64.50 | 37.31 | 37.35 | 36.30 | 28.79 | 40.85 |
| GPT-4 | generation | N/A | **88.00** | 37.31 | 56.52 | 79.79 | 64.13 | 65.15 |
| | | prompt-calib1 | 83.50 | **45.52** | **65.21** | 84.04 | 70.65 | **69.78** |
| | | prompt-calib2 | 83.00 | 45.52 | 64.13 | 84.04 | 69.02 | 69.14 |
| | | prompt-calib3 | 82.00 | 44.40 | 61.96 | **86.17** | **73.91** | 69.69 |
| Vicuna-7B | generation | vanilla | **71.50** | 25.00 | 29.35 | **48.94** | 23.91 | 39.74 |
| | | contrast (inst) | 70.00 | 61.57 | 35.87 | 42.55 | 64.13 | 54.82 |
| | | contrast (answer) | 69.00 | **65.67** | **39.13** | 39.36 | **72.83** | **57.20** |

Table 1: Experiment results of different bias mitigation methods on LLMBar

### 4.3 Compromise between Bias and Accuracy

In this part, we would like to study the correlation between bias mitigation and evaluation accuracy. As we discussed before, while current LLM evaluators have been proved to be biased towards the superficial quality such as verbosity, formality, etc, these qualities are also important attributes of human preference and should not be excluded. We control the extent of bias mitigation by adjusting the coefficient factor $\alpha$ in Equation 1, and the change of accuracy with respect to $\alpha$ is shown in Figure 4.

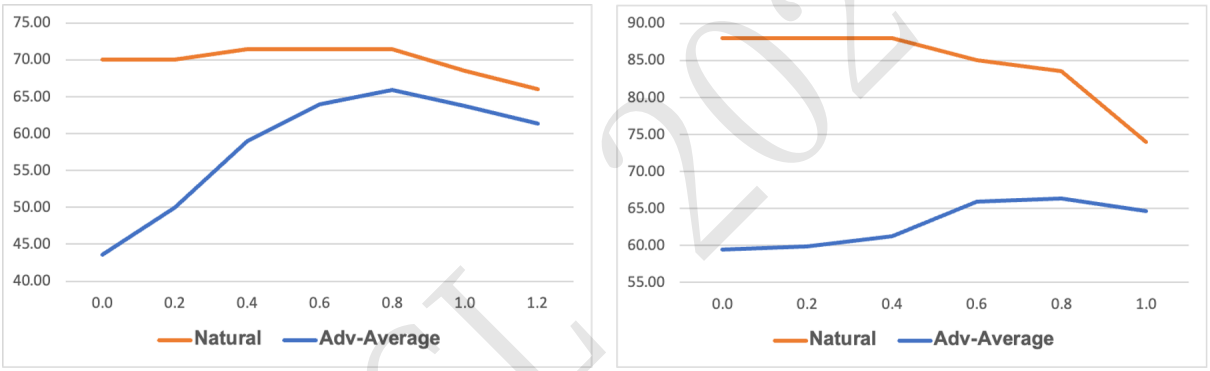



Figure 4: The variation of accuracy with respect to different coefficient. Left denotes the result of Text-davinci-003 on probability calibration, and right denotes the result of GPT-4 on prompt calibration.

As we can see, at the beginning of bias mitigation, the accuracy on the Natural set keeps stable, while the accuracy on the Adversarial set increase gradually. But after the factor is larger than 0.8, the accuracy on both Natural set and Adversarial set would decrease. This verifies our hypothesis that there exists a compromise between bias mitigation and evaluation bias. Bias mitigation, to a certain extend, is helpful for improving the evaluation accuracy, but too much mitigation would lead to performance degradation eventually, and the superficial quality should not be excluded from the prediction process.

### 4.4 Negative Sampling for Contrastive Training

For contrastive training, the most important factor is the construction of negative samples. The negative samples should hold better superficial quality, while maintaining mismatch with the instruction. There is a compromise that lies here: The negative sample should not deviate from the instruction too much to improve the difficulty of contrastive training, neither should it be too close to the instruction to keep the negativity. Moreover, we also want to ensure the negative sample always holds better superficial quality. To achieve the most effective negative sampling, we conducted the following experiments:

1. Adjacent Instruction: Select the most similar instruction within a scale;

| Model | Method | Mitigation | Natural | Neighbor | Manual | GPTOut | GPTInst | Average |
|---|---|---|---|---|---|---|---|---|
| Vicuna-7B | generation | Vanilla | **71.50** | 25.00 | 29.35 | **48.94** | 23.91 | 39.74 |
| | | Adjacent Instruction | 70.00 | 61.57 | 35.87 | 42.55 | 64.13 | 54.82 |
| | | + Sup-Scorer (Max) | 66.00 | 45.52 | 30.43 | 42.55 | 51.09 | 47.12 |
| | | + Sup-Scorer (Avg) | 66.00 | 61.94 | 31.52 | 39.36 | 71.74 | 54.11 |
| | | + Adjacent Answer | 69.00 | **65.67** | **39.13** | 39.36 | **72.83** | **57.20** |

Table 2: Experiment results of different negative sampling methods for contrastive training.

2. Adjacent Instruction + Adjacent Answer: Similar to 1, but this time directly select the answer that is most similar to the correct answer in the dataset within a scale, and add the data of the Adjacent Instruction;

3. Adjacent Instruction + Superficial Scorer (Max): Train a classifier to discern the samples with better superficial quality. To achieve this goal, we train a binary classifier with only the target input, therefore the obtained classifier can only predict the evaluation based on the target input. After that, the classifier is used to select the sample with better superficial quality within the scale;

4. Adjacent Instruction + Superficial Scorer (Avg): Similar to 3, but the score of superficial quality and semantic similarity is averaged for sampling;

As shown in Table 2, the superficial scorer does not introduce improvement. This notifies us that the superficial quality can not be simply attributed solely to the target side. While target-only classifier being verified as an effective bias measurement for traditional translation evaluation, the situation for LLM evaluation is more complicated.

On the other hand, the adjacent selection based both on the instruction and the answer achieves better results. This suggests the instruction similarity modeling and answer similarity modeling obtain information of the original data from different perspectives, which promotes the training of the judge model. The instruction similarity can help understand the nuances of the tasks, while the answer similarity can help understand the solution approaches. Therefore, it is better to perform negative sampling both on the instructions and answers.

## 5 Conclusion

In this work, we propose two methods for mitigating the bias of LLM-as-a-Judge. For closed-source judge models, we propose to mitigate the bias by calibration. For open-source judge models, we propose to mitigate the bias by pairwise contrastive training. We apply our methods on the LLMBar benchmark, effectively mitigating the bias of LLM evaluation while maintaining a comparable performance.

As the LLM-as-a-Judge becomes a paradigm for evaluating the LLM performance, it is advisable to exert more alert on the bias of LLM-based evaluators. In the future, we will dedicate more effort to the bias mitigation of LLM-as-a-Judge into more fine-grained categories. We will also conduct research about the inference bias on other LLM-based applications.

## Acknowledgements

This work is supported by National Natural Science Foundation of China (62376075, 62276077, U1908216, 62376076), Key R&D Program of Yunnan (202203AA080004) and Shenzhen College Stability Support Plan (No. GXWD20220811170358002).

## References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In Jong C. Park, Yuki Arase, Baotian Hu, Wei Lu, Derry Wijaya, Ayu Purwarianti, and Adila Alfa Krisnadhi, editors, *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali, November. Association for Computational Linguistics.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection.

Hui Huang, Yingqi Qu, Jing Liu, Muyun Yang, and Tiejun Zhao. 2024. An empirical study of llm-as-a-judge for llm evaluation: Fine-tuned judge models are task-specific classifiers.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacaeval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.

Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. 2023. Verbosity bias in preference labeling by large language models. *arXiv preprint arXiv:2310.10076*.

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2023. One embedder, any task: Instruction-finetuned text embeddings. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1102–1121, Toronto, Canada, July. Association for Computational Linguistics.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large language models are not fair evaluators.

Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. 2024. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization.

Jifan Yu, Xiaozhi Wang, Shangqing Tu, Shulin Cao, Daniel Zhang-Li, Xin Lv, Hao Peng, Zijun Yao, Xiaohan Zhang, Hanming Li, Chunyang Li, Zheyuan Zhang, Yushi Bai, Yantao Liu, Amy Xin, Kaifeng Yun, Linlu GONG, Nianyi Lin, Jianhui Chen, Zhili Wu, Yunjia Qi, Weikai Li, Yong Guan, Kaisheng Zeng, Ji Qi, Hailong Jin, Jinxin Liu, Yu Gu, Yuan Yao, Ning Ding, Lei Hou, Zhiyuan Liu, Xu Bin, Jie Tang, and Juanzi Li. 2024. KoLA: Carefully benchmarking world knowledge of large language models. In *The Twelfth International Conference on Learning Representations*.

CCL 2024

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. 2023. Evaluating large language models at evaluating instruction following. *arXiv preprint arXiv:2310.07641*.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2023. Judgelm: Fine-tuned large language models are scalable judges. *arXiv preprint arXiv:2310.17631*.

CCL 2024

# PPDAC: A Plug-and-Play Data Augmentation Component for Few-shot Extractive Question Answering

**Qi Huang**[1,2] **Han Fu**[2] **Wenbin Luo**[1] **Mingwen Wang**[1,2] **Kaiwei Luo**[2]
[1] School of Computer and Information Engineering, Jiangxi Normal University,
Nanchang, Jiangxi 330022, China
[2]School of Digital Industry, Jiangxi Normal University,
Shangrao, Jiangxi 334000, China
{huangqi, 202141600077, lwb, mwwang, luokaiwei}@jxnu.edu.cn

## Abstract

Extractive Question Answering (EQA) in the few-shot learning scenario is one of the most challenging tasks of Machine Reading Comprehension (MRC). Some previous works employ external knowledge for data augmentation to improve the performance of few-shot extractive question answering. However, there are not always available external knowledge or language- and domain-specific NLP tools to deal with external knowledge such as part-of-speech taggers, syntactic parsers, and named-entity recognizers. In this paper, we present a novel Plug-and-Play Data Augmentation Component (PPDAC) for the few-shot extractive question answering, which includes a paraphrase generator and a paraphrase selector. Specifically, we generate multiple paraphrases of the question in the (question, passage, answer) triples using the paraphrase generator and then obtain highly similar statements via paraphrase selector to form more training data for fine-tuning. Extensive experiments on multiple EQA datasets show that our proposed plug-and-play data augmentation component significantly improves question-answering performance, and consistently outperforms state-of-the-art approaches in few-shot settings by a large margin.

## 1 Introduction

Extractive Question Answering (EQA) presents a significant challenge within the domain of machine reading comprehension, requiring a nuanced understanding of both questions and passages to accurately identify the text fragments that serve as answers. High-performance EQA algorithms are essential for enabling retrieval systems to swiftly locate relevant passages and text fragments based on user queries. However, the reality is that not all scenarios offer a wealth of labeled samples, leading to diminished effectiveness of current EQA algorithms. This gap underscores the importance of investigating EQA algorithms within few-shot learning contexts, which are not only more aligned with real-world situations but also more feasible for practical implementation.

The primary issue confronting existing EQA algorithms in few-shot scenarios is performance degradation due to a disconnect between the pre-training objectives of Masked Language Modeling (MLM) and the fine-tuning objectives of downstream Machine Reading Comprehension (MRC) tasks. While (Glass et al., 2020; Ram et al., 2021; Chada and Natarajan, 2021) have attempted to bridge this gap by integrating downstream tasks into pre-training or by aligning the fine-tuning framework with the pre-training framework (known as the prompt-based fine-tuning paradigm, or prompt-tuning), these approaches often require extensive computational resources and overlook the potential of external knowledge bases for data augmentation in few-shot EQA.

In response to these challenges, another avenue of research has focused on leveraging external knowledge to either generate training data or enhance the representation of prompt-tuning paradigm methods. Specifically, (Lewis et al., 2019) generate context, question, and answer triples via the traditional NLP pipeline from the external knowledge base, which selects noun phrases and named entities in Wikipedia paragraphs as potential answers, and then masks from the context to create pseudo-questions. Moreover, (Wang et al., 2022) proposed a novel framework named knowledge enhanced contrastive prompt-tuning,

©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1320-1333, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

which enhances the representation of the question by the support of the rich semantics from the external knowledge base and the passage context. Despite the potential of these approaches, they face significant hurdles, including limited availability of external knowledge or language- and domain-specific NLP tools, the risk of error accumulation in the NLP pipeline, and a scarcity of data augmentation methods specifically designed for few-shot EQA tasks.

To address these issues, this paper introduces a novel Plug-and-Play Data Augmentation Component (PPDAC), aimed at generating paraphrases of questions within (question, passage, answer) triples to enrich the training data. The PPDAC comprises a paraphrase generator, a paraphrasing filter, and a paraphrase selector, which work together to transform the original question statements into multiple new statements with similar semantics but varied syntactic structures. This approach substantially increases the volume of training data available for fine-tuning. Our experimental evaluation across various EQA baselines in few-shot scenarios reveals significant performance improvements, showcasing the effectiveness of PPDAC in addressing the challenges of few-shot EQA without relying on external knowledge bases.

In summary, our contributions can be summarized as follows:

- We present a novel Plug-and-Play Data Augmentation Component (PPDAC) for data augmentation in few-shot EQA tasks.
- PPDAC replaces the external knowledge base by converting the statements of the questions in the data (questions, paragraphs, answers) into multiple new statements and selecting highly similar statements to form more training data for fine-tuning, thereby improving the applicability of data enhancement methods.
- Experimental results show that our proposed PPDAC can effectively improve the performance of few-shot EQA algorithms.

## 2 Related Work

Extractive Question Answering (EQA) methods based on pre-trained language models resulted in performance close to (and sometimes exceeding) a human performance when fine-tuned on several QA benchmarks(Kenton and Toutanova, 2019; Brown et al., 2020; Bao et al., 2020; Raffel et al., 2020). These methods need to be fine-tuned on tens of thousands of examples to obtain this result. However, Only a small amount of annotated training data is available in a more realistic and practical scenario, which results in their performance degrading significantly. To address this issue, researchers started from two aspects: 1) reducing the difference between the pre-training objective of Masked Language Modeling (MLM) and the fine-tuning objective of downstream MRC tasks; 2) utilizing external knowledge to generate the training data or enhancing the representation.

To reduce the difference between the pre-training objective and the fine-tuning objective, some researchers construct new pre-training tasks specifically tailored to the EQA task, and other researchers propose a new fine-tuning framework that is directly aligned with the pre-training framework. For instance, (Ram et al., 2021) designed a new pretraining scheme tailored for few-shot question answering, i.e. using a recurring span selection as the objective of pre-training models. (Castel et al., 2021) presents a decoding algorithm on the fine-tuning phase that efficiently finds the most probable answer span in the passage. And (Chada and Natarajan, 2021) explore a simple fine-tuning framework that leverages pre-trained text-to-text models and is directly aligned with the pre-training framework, in terms of both the input-output design and the training objective. However, these methods may cost a lot of computational resources during the pre-training phase and ignore the external knowledge base for data augmentation in few-shot question answering.

Therefore, some researchers explore another branch of the few-shot EQA that studies the data augmentation methods to generate the training data or enhance the representation from the external knowledge base. (Lewis et al., 2019; Glass et al., 2020) exploit the traditional NLP pipeline to select noun phrases and named entities in Wikipedia paragraphs as potential answers, and then mask these potential answers from the context to create pseudo-questions. The difference is that (Lewis et al., 2019)

CCL 2024

utilizes unsupervised machine translation methods to translate these pseudo-questions into real ones, while (Glass et al., 2020) exploits information retrieval to find new text passages that can answer the pseudo-questions. Moreover, (Wang et al., 2022) presents a novel framework named knowledge enhanced contrastive prompt-tuning, which enhances the representation of the question by the support of the rich semantics from the external knowledge base and the passage context. However, the above studies assume access to external knowledge or language- and domain-specific NLP tools such as part-of-speech taggers, syntactic parsers, and named-entity recognizers, which may not always be available.

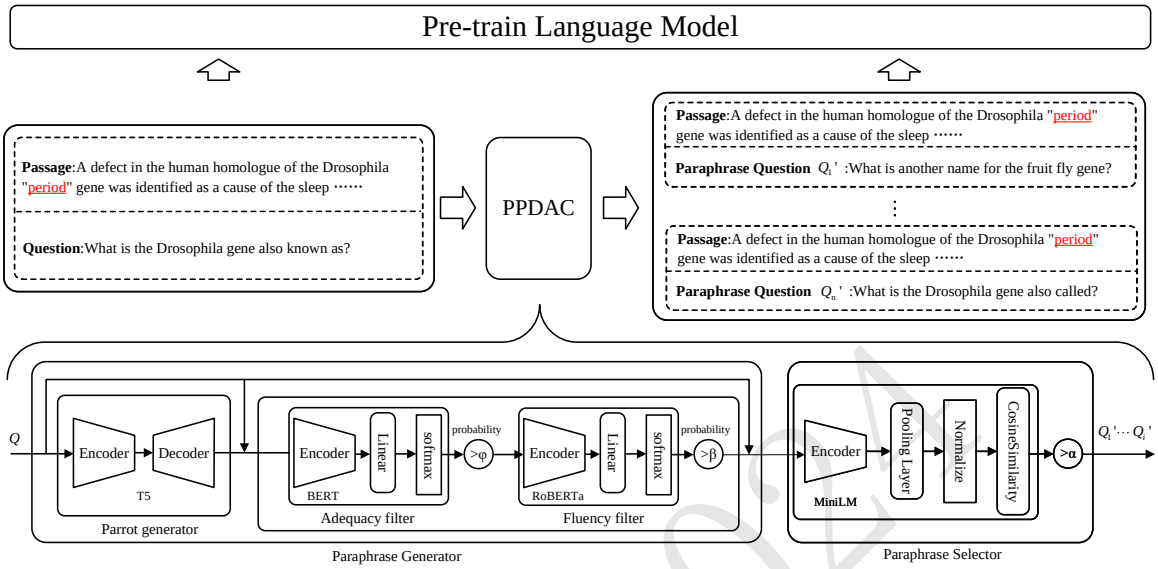



Figure 1: The overview of the PLM-based framework for few-shot EQA with our proposed component PPDAC. Given a (question, passage, answer) triples, our proposed PPDAC converts the statement of the question in (question, passage, answer) triples into multiple new statements by the paraphrase generator, and then obtains highly similar statements via paraphrase selector to form more training data for fine-tuning the pre-train language model.

## 3 The Proposed Plug-and-play Component

In this section, we formally introduce the few-shot EQA task and our proposed plug-and-play component in detail. The overview of the PLM-based framework for few-shot EQA with our proposed component is shown in Figure 1.

### 3.1 Task Statement

Given a passage $P = (p_1, \ldots, p_n)$ and the corresponding question $Q = (q_1, \ldots, q_m)$, the goal of EQA is to find a sub-string of the passage as the answer $A = (p_k, \ldots, p_l)$, where n and m represent the lengths of the passage and the question, respectively. Here, $p_i (i = 1, \ldots, n)$ and $q_j (j = 1, \ldots, m)$ denote the tokens in $P$ and $Q$, respectively. $k$ and $l$ refer to the start and end position of the answer, restricted to $[1, n]$. The conventional EQA methods based on language models need to be finetuned on tens of thousands of samples to achieve satisfactory performance. However, only a few samples $(P, Q, A)$ can fine-tune the pre-trained language models in the few-shot scenarios, i,e. few-shot EQA task. Rather than the methods exploiting external knowledge for few-shot EQA, we explore a plug-and-play data augmentation component to generate training data for fine-tuning the pre-trained language models. In the following, we will introduce the details of our proposed plug-and-play data augmentation component, which consists of a paraphrase generator and a paraphrase selector. Subsequently, we provide a comprehensive exposition of these individual components.

CCL 2024

### 3.2 Paraphrase Generator

The Paraphrase Generator is the cornerstone of our proposed component inspired by (Damodaran, 2021), which aims to convert the question of (question, passage, answer) triples into a series of new statements and preserve the core meaning of the question. This generator encompasses three essential components: parrot generator, adequacy filter, and fluency filter. The parrot generator utilizes a T5 model with a beam search strategy to generate various statements of question. Then the adequacy filter calculates the adequacy score between the generated statements and the original question via a pre-trained model to filter out the statements below the threshold. Finally, the fluency filter exploits a pre-trained model to evaluate the fluency score of the statements and filter out the statements below the threshold.

#### 3.2.1 Parrot Generator

To generate the paraphrase of the question in (question, passage, answer) triples, the parrot generator utilizes a transformer-based T5 model for the task of autoregressive language generation. the parrot generator exploits the encoder of T5 model to transform the given question Q into a representation X containing contextual semantic information. The formalization process is as follows:

$$X_Q = Encoder_{T5}(Q) \tag{1}$$

Next, according to the contextual semantic representation $X_Q$, the parrot generator uses the decoder of the T5 model to decode the semantic representation to obtain a paraphrase of the original question. To enhance the diversity of the generated paraphrase, we introduce the beam search algorithm to decoding. Specifically, during the decoding process, the beam search algorithm preserves the num beam most likely probable, that is a beam represents a word, in each time step. Then when the search ends, the beam sequence or say the word sequence with the highest cumulative probability is selected as the output of the algorithm. The above process increases the likelihood of generating sequences of high-probability word combinations. However, the sentence generated by the beam search algorithm still contains repetitive sequences of words.

To this end, inspired by the n-grams penalty introduced in (Paulus et al., 2018; Klein et al., 2017), we utilize a diversity penalty parameter to control the recurrence of words. The diversity penalty parameter is determined by comparing the currently generated token with a beam from another group at a specific time. If there is a match between the generated token and the prior beam then remove the token from the current beam. In summary, the above operations ensure that the parrot generator can generate sentences that are fluent and have as few repeated words as possible.

#### 3.2.2 Paraphrase Filter

In our approach, we enhance the paraphrase generation process by incorporating two critical filters: an adequacy filter and a fluency filter. Both filters leverage finely-tuned pre-trained models to ensure the high quality of generated paraphrases, aligning them closely with the semantic intent and linguistic standards of the original questions.

**Adequacy Filter** The adequacy filter employs the parrot_adequacy_model, a specialized textual entailment classifier derived from fine-tuning the pre-trained RoBERTa model. This model adeptly identifies whether the generated paraphrase maintains the same semantic essence as the original question, categorizing relationships as entailment, contradiction, or neutral. We utilize this classifier to assess the neutrality of paraphrases, discarding those that do not meet a predefined neutrality threshold. This ensures that only semantically consistent paraphrases are retained [0].

**Fluency Filter** Simultaneously, the fluency filter applies the parrot_fluency_model, a binary classifier fine-tuned from the pre-trained BERT model, designed to evaluate the linguistic fluency of paraphrases. By quantifying the fluency score, this model filters out paraphrases that fall below a certain fluency threshold, ensuring that the final selections are not only semantically accurate but also linguistically polished [1].

[0]The $parrot\_adequacy\_model$ is available for community use at https://huggingface.co/prithivida/parrot_adequacy_model.
[1]The $parrot\_fluency\_model$ is available for community use at https://huggingface.co/prithivida/parrot_fluency_model.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1320-1333, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

By integrating these filters, our method significantly enhances the quality of paraphrases used in training, directly contributing to the improved performance of our few-shot EQA system. The open-source availability of both models underscores our commitment to transparency and community collaboration.

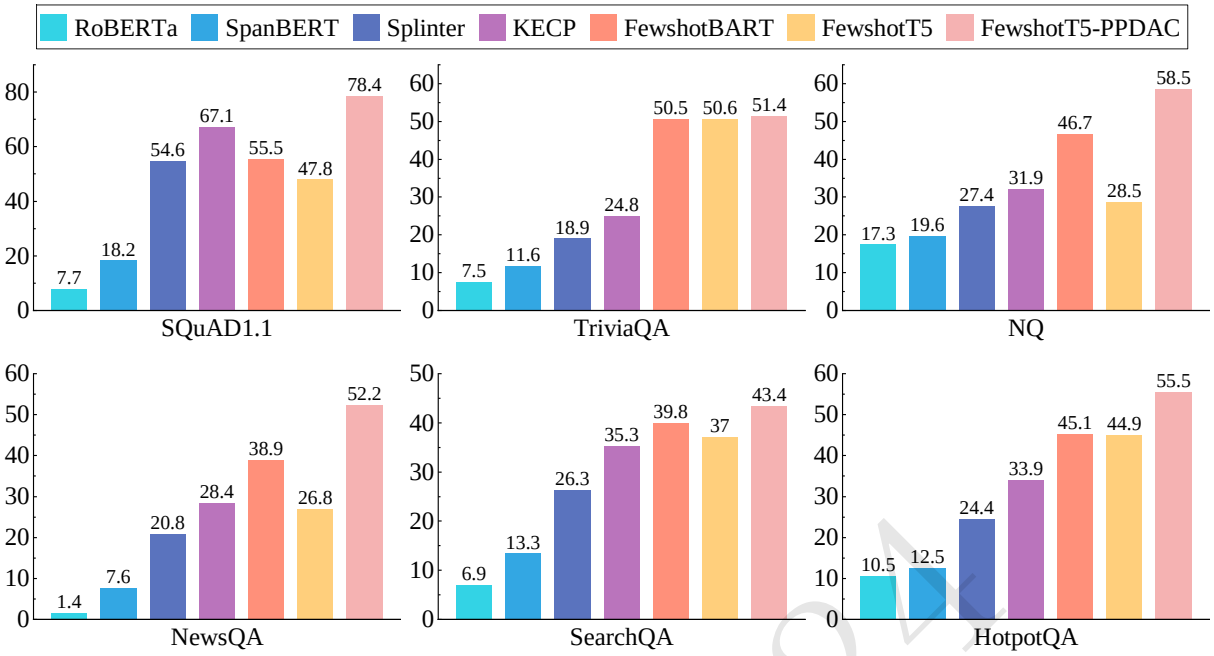



Figure 2: The performance of different methods over 6 datasets in the 16-shot setting, where FewshotT5-PPDAC is the FewshotT5-based model with our PPDAC. The value of performance is the mean of the F1 value across 5 randomly sampled training sets of the same size.

### 3.3 Paraphrase Selector

To obtain a high-quality paraphrase of the question for fine-tuning the pre-trained language model, we designed a paraphrase selector to remove redundancy and enhance the diversity of the question statement. The paraphrase selector assesses the diversity and quality of generated questions by calculating the similarity between the generated questions and the original questions. The encoding of both the generated and original questions is conducted utilizing the pre-trained MiniLM. The formula of similarity is as follows:

$$X_Q = Encoder_{MiniLM}(Q) \tag{2}$$

$$X_{Q'_i} = Encoder_{MiniLM}(Q'_i) \tag{3}$$

$$\cos_{Q,Q'_i} = \frac{X_Q \bullet X_{Q'_i}}{\sqrt{X_Q{}^2} \times \sqrt{X^2_{Q'_i}}} \tag{4}$$

where $Q'_i$ represents the $i$ th question generated by the paraphrase generator, i=[1,...,n]. Then we select the generate questions $Q'$ whose value of $cos_{Q,Q'_i}$ is greater than $\alpha$ to form the ($Q'$, P, A) for fine-tuning, $\alpha$ is our defined threshold of similarity.

As we all know, the value of the $\alpha$ is the key factor to maintaining a dynamic balance between the diversity and quality of new questions generated. Therefore the selection of $\alpha$ value is essential for our proposed component.

Although both the Paraphrase Filter and the Paraphrase Selector are involved in filtering high-quality questions, they serve different yet complementary roles within the augmentation pipeline. The Paraphrase Filter acts as the first gatekeeper, ensuring semantic consistency and fluency. The Paraphrase

| | Dataset | Train | Dev | Test |
|---|---|---|---|---|
| split I | SQuAD | 86,588 | 10,507 | - |
| | NewsQA | 74,160 | 4,212 | - |
| | TriviaQA | 61,688 | 7,785 | - |
| | SearchQA | 117,384 | 16,980 | - |
| | HotpotQA | 72,928 | 5,904 | - |
| | NQ | 104,071 | 12,836 | - |
| split II | BioASQ | - | 1,504 | 1,518 |
| | TextbookQA | - | 1,503 | 1,508 |

Notes: the test data of the Split II in the MRQA 2019 shared task is not available.

Table 1: The statistics of the two subsets of the MRQA 2019 shared task.

Selector further optimizes the output by enhancing the overall quality of the questions. This division of responsibilities prevents redundancy and ensures that the generated questions are both accurate and natural, thereby improving the quality of the augmented data.

## 4 Experiments

In this section, we conduct extensive experiments to verify the improvement of the proposed plug-and-play component PPDAC for the few-shot EQA methods[2].

### 4.1 Baselines

To thoroughly evaluate the enhancement brought by our proposed component to few-shot EQA methods, we benchmark against a broad spectrum of methods, including those introduced beyond 2019 to ensure our comparison encapsulates the latest advancements in the field:

1) **RoBERTa** (Liu et al., 2019): A pre-trained model that refines BERT by implementing a dynamic masking strategy, establishing a strong baseline for numerous NLP tasks.

2) **SpanBERT** (Joshi et al., 2020): A variant of BERT that focuses on predicting spans of text, rather than individual tokens, for improved performance on span-based tasks like EQA.

3) **Splinter** (Ram et al., 2021): Pioneers the use of recurring span selection as a pre-training task specifically tailored for enhancing EQA performance.

4) **FewshotBART** and **FewshotT5** (Chada and Natarajan, 2021; Raffel et al., 2020; Lewis et al., 2020): Leverage the respective strengths of BART and T5 within an aligned fine-tuning framework, showcasing their adaptability to few-shot learning contexts.

5) **FewshotBARTL** (Chada and Natarajan, 2021; Lewis et al., 2020): Within the identical fine-tuning framework, the BART-large model possesses a parameter count significantly exceeding that of both BART-base and T5-base models. Consequently, its outcomes are utilized merely for comparative reference.

6) **KECP** (Wang et al., 2022): A method that transforms the EQA task into a non-autoregressive Masked Language Modeling generation problem, enhances the representation of the question by the external knowledge base and trains the model using the MLM and contrastive learning objectives.

### 4.2 Datasets

We evaluate our proposed component on a subset of the MRQA 2019 shared task (Fisch et al., 2019), which contains extractive question-answering datasets in a unified format and the answer is a single span in the given text passage. The MRQA 2019 shared task is split into two subsets (i.e. Split I and Split II) and the statistics of each dataset on two subsets are shown in Table 1.

**Split I of the MRQA 2019 shared task**: It is a shared task consisting of 6 EQA datasets: SQuAD1.1 (Rajpurkar et al., 2016), NewsQA (Trischler et al., 2017), TriviaQA (Joshi et al., 2017), SearchQA (Dunn et al., 2017), HotpotQA (Yang et al., 2018), and Natural Question (Kwiatkowski et al., 2019). Following previous works (Ram et al., 2021; Wang et al., 2022), we utilize the official development set for evaluation, as the testing set is not publicly available.

[2]The source codes and datasets used will be available on the GitHub repository after being accepted.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1320-1333, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

| Model | SQuAD | TriviaQA | NQ | NewsQA | SearchQA | HotpotQA | BioASQ | TextbookQA |
|---|---|---|---|---|---|---|---|---|
| *16 Examples* | | | | | | | | |
| Splinter | 54.6±6.4 | 18.9±4.1 | 27.4±4.6 | 20.8±2.7 | 26.3±3.9 | 24.0±5.0 | 28.2±4.9 | 19.4±4.6 |
| Splinter-PPDAC | **58.2±2.5** | **21.4±4.8** | 26.9±2.2 | 19.3±2.2 | **28.7±2.1** | **28.8±2.6** | **28.8±1.0** | 19.4±3.9 |
| FewshotBART | 55.5±2.0 | 50.5±1.0 | 46.7±2.3 | 38.9±0.7 | 39.8±0.1 | 45.1±1.5 | 49.4±0.1 | 19.9±1.2 |
| FewshotBART-PPDAC | **57.4±0.8** | 47.1±0.1 | **47.7±0.3** | 37.8±0.6 | **45.9±1.2** | **60.4±1.1** | **49.9±1.8** | **20.1±1.0** |
| FewshotT5 | 47.8±6.9 | 50.6±4.9 | 28.5±14.5 | 26.8±2.7 | 37.0±3.3 | 44.9±3.5 | 46.3±5.9 | 25.9±5.0 |
| FewshotT5-PPDAC | **78.4±0.2** | **51.4±2.0** | **58.5±1.0** | **52.2±0.5** | **43.4±2.1** | **55.5±0.3** | **58.6±0.4** | **32.1±2.5** |
| *128 Examples* | | | | | | | | |
| Splinter | 72.7±1.0 | 44.7±3.9 | 46.3±0.8 | 43.5±1.3 | 47.2±3.5 | 54.7±1.4 | 63.2±4.1 | 42.6±2.5 |
| Splinter-PPDAC | **73.8±0.7** | **45.1±1.0** | **48.8±0.4** | **43.6±0.9** | **48.0±2.1** | **55.3±0.5** | **64.4±2.0** | 37.9±2.0 |
| FewshotBART | 68.0±0.3 | 50.1±1.8 | 53.9±0.9 | 47.9±1.2 | 58.1±1.4 | 54.8±0.8 | 68.5±1.0 | 29.7±2.4 |
| FewshotBART-PPDAC | **71.6±1.0** | **53.9±1.2** | **54.1±0.2** | **49.9±0.1** | **59.1±0.2** | **67.1±0.6** | **72.4±0.2** | **37.8±0.4** |
| FewshotT5 | 64.6±6.1 | 51.7±3.1 | 47.0±4.6 | 40.0±1.9 | 57.0±4.5 | 56.1±3.7 | 68.2±3.6 | 33.6±2.1 |
| FewshotT5-PPDAC | **86.3±0.1** | **52.8±0.3** | **64.3±0.1** | **60.9±0.4** | **60.8±0.4** | **64.4±0.9** | **73.9±0.9** | **51.5±0.8** |
| *1024 Examples* | | | | | | | | |
| Splinter | 82.8±0.8 | 64.8±0.9 | 65.5±0.5 | 57.3±0.8 | 67.3±1.3 | 70.3±0.8 | 91.0±1.0 | 54.5±1.5 |
| Splinter-PPDAC | 82.8±0.7 | **64.9±1.0** | **66.2±0.4** | 57.0±0.9 | 67.3±1.1 | **70.4±0.5** | 91.0±0.5 | **54.7±1.3** |
| FewshotBART | 79.6±0.1 | 55.5±0.7 | 61.4±0.1 | 58.9±0.6 | 70.0±0.1 | 65.8±0.7 | 92.2±0.8 | 50.0±0.2 |
| FewshotBART-PPDAC | **79.7±0.1** | 55.5±0.9 | **62.4±0.1** | 58.7±0.1 | **70.5±0.2** | 65.0±0.6 | **92.5±0.2** | **51.8±0.4** |
| FewshotT5 | 88.6±0.2 | 60.0±0.5 | 69.4±0.3 | 64.7±0.2 | 73.0±0.2 | 70.2±0.2 | 86.9±0.2 | 59.7±0.5 |
| FewshotT5-PPDAC | **89.0±0.1** | **61.0±0.1** | 69.4±0.1 | **64.9±0.1** | **73.1±0.1** | **71.2±0.2** | **87.8±0.1** | **60.1±0.6** |

Notes: this is a part of the whole experimental results, the whole please consult the Appendix

Table 2: The gain results of the PPDAC components are compared with the baseline results -PPDAC indicates the baseline with our PPDAC. The standard deviation is calculated across 5 runs with different seeds. NQ stands for Natural Questions dataset. The table highlights improvements in effectiveness with bolded numbers, while the best results achieved with the same dataset and number of training samples are denoted in blue.

**Split II of the MRQA 2019 share task**: Following (Ram et al., 2021), we also select two datasets from Split II of the MRQA 2019 shared task that were annotated by domain experts to evaluate our proposed component: namely BioASQ (Tsatsaronis et al., 2015) and TextbookQA (Kembhavi et al., 2017). These datasets only included publicly available development sets within the MRQA platform, each containing approximately 1,500 examples. For evaluation, we sample 400 examples from each dataset to create a separate test set, and the remaining data serves as the training data.

### 4.3 Implementation Details

Our experimental framework adopts a methodology similar to (Ram et al., 2021) for constructing few-shot training and development sets across various EQA datasets. We achieve this by randomly selecting K samples from the original training sets, where K ranges from 16 to 1024. Due to the unavailability of the original test sets, evaluations are conducted on the complete development sets.

Baseline models are selected based on pre-trained language models (PLMs) available through HuggingFace, with initial hyperparameters set accordingly. Model training employs the Adam optimizer(Kingma, 2014) with bias-corrected moment estimates, as recommended for few-shot learning contexts(Zhang et al., 2021). The training learning rate is fixed at 2e-5, incorporating a warm-up phase covering the initial 10% of steps, succeeded by a linear decay. Training durations are set to 40 epochs or 200 steps, whichever is greater, for experiments involving up to 1024 examples. The batch size is configured at 8 for training with 128 or fewer examples and reduced to 4 for 1024-shot learning scenarios.

For the paraphrase selector, the threshold $\alpha$is set to 0.9 based on the experience of a large number of experiments. To ensure the robustness of our findings, experiments are conducted with five distinct random seeds $\{42, 43, 44, 45, 46\}$, and performance metrics are averaged. The evaluation metric employed is the F1 score, measuring the average overlap between predicted and true answer texts, following the protocol established by (Ram et al., 2021). Our models are implemented using PyTorch and trained on an RTX 3090 GPU server, ensuring computational efficiency and reproducibility of results.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1320-1333, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

### 4.4 Few-shot EQA performance

Our comprehensive evaluation showcases the significant enhancements brought about by the Plug-and-Play Data Augmentation Component (PPDAC) in few-shot Extractive Question Answering (EQA) settings. Figure 2 vividly illustrates the superior performance of FewshotT5-PPDAC, our method that leverages PPDAC for data augmentation, against all baseline models in an experimental setup comprising only 16 training examples. In this rigorous test, FewshotT5-PPDAC attains remarkable F1 scores of 78.4%, 51.4%, 58.5%, 52.2%, 43.4%, and 55.5% across six distinct datasets within Split I of the MRQA 2019 shared task. These scores represent substantial improvements over the leading baselines by margins of 11.3%, 0.8%, 11.8%, 13.3%, 3.6%, and 10.4%, respectively, underscoring the efficacy of PPDAC in amplifying question-answering performance in constrained data environments. Notably, when compared against the advanced KECP method in Figure 2, our approach demonstrates a commanding lead of 11.3%, 26.6%, 26.6%, 23.8%, 8.1%, and 21.6% across the six datasets, respectively. While our primary goal was to apply our proposed components within the KECP framework to validate their effectiveness, the lack of open-source availability for KECP constrained us to using it as a baseline for performance comparisons only.

To extend our investigation into the PPDAC's impact on baseline models, we integrated our component across a diverse array of baselines, observing performance across varying training sample sizes. The outcomes, detailed in Table 2, reveal a consistent trend of performance enhancement with PPDAC, particularly pronounced in smaller sample sizes. It's important to note that Table 2 showcases only a subset of our experimental data for clarity and brevity. The comprehensive dataset, encompassing all baseline models and the entire range of training sample sizes, is extensive and therefore has been included in the appendix for those interested in a more detailed exploration. This approach ensures that readers can grasp the immediate benefits of PPDAC from the main text while providing access to the full breadth of our findings for deeper analysis.

This observation is critical, highlighting PPDAC's role in not only boosting EQA performance but also in reducing the variability of model outcomes, as evidenced by a decreased standard deviation among the enhanced baselines. Notably, our analysis identified instances where performance gains were less marked, particularly in datasets characterized by complex questions that exhibit significant syntactic or lexical variance from their answers. This suggests a potential limitation of data augmentation strategies in handling the intricate semantics of complex questions, a factor that merits further exploration to optimize the efficacy of augmentation techniques in such contexts.

A pivotal finding of our study is the pronounced benefit of PPDAC in scenarios with limited data, a testament to its utility in enhancing the robustness and reliability of EQA models under the few-shot paradigm. While the performance improvements are more modest in settings with larger data volumes, the reduction in standard deviation after the application of the PPDAC plugin across most baseline models highlights an increased stability and consistency in model performance.

In summary, our research underscores the transformative potential of PPDAC in bolstering the capabilities of few-shot EQA methods. By facilitating more effective data augmentation, PPDAC not only enhances model performance across a spectrum of training conditions but also contributes to a deeper understanding of the dynamics at play in few-shot learning environments. These insights pave the way for future advancements in EQA, particularly in the development of more resilient and adaptable models suited to the diverse challenges inherent in real-world applications.

### 4.5 The Efficiency of PPDAC

In addition to our comprehensive performance evaluation, we also assessed the time and space efficiency of the PPDAC to demonstrate its practicality in real-world applications. Given that PPDAC operates as a pre-processing step prior to the fine-tuning of pre-trained models, it introduces minimal overhead, ensuring that the efficiency of the fine-tuning process remains largely unaffected. This is a crucial advantage, as it allows for the significant enhancements in EQA performance without imposing substantial additional computational costs.

To quantify the impact of PPDAC, we measured the additional time and space required for data aug-

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1320-1333, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

| #Train Samples | Training time | Evaluation time |
|---|---|---|
| Splinter | 25s | 46.78s |
| Splinter-PPDAC | 64s | 47.37s |
| FewshotBART | 48s | 148.3s |
| FewshotBART-PPDAC | 123s | 151.2s |
| FewshotT5 | 82s | 124.2s |
| FewshotT5-PPDAC | 230s | 125.6s |

Table 3: The table presents a comprehensive comparison of the training and testing times for three models: Splinter, Fewshot-BART, and Fewshot-T5, both before and after the application of the PPDAC component.

| #Train Samples | 16 | 32 | 128 |
|---|---|---|---|
| **Splinter-PPDAC** | **58.2%** | **64.3%** | **73.8%** |
| Splinter-PPDAC(w/o. PF) | 56.4% | 62.2% | 72.9% |
| Splinter-PPDAC(w/o. PS) | 56.1% | 60.6% | 73.2% |
| Splinter-PPDAC(w/o. PF & PS) | 55.6% | 60.1% | 72.7% |
| Splinter | 54.6% | 59.2% | 72.7% |
| **FewshotBART-PPDAC** | **57.4%** | **60.4%** | **71.6%** |
| FewshotBART-PPDAC(w/o. PF) | 56.2% | 57.9% | 68.9% |
| FewshotBART-PPDAC(w/o. PS) | 56.6% | 58.2% | 69.1% |
| FewshotBART-PPDAC(w/o. PF & PS) | 56.0% | 57.3% | 68.6% |
| FewshotBART | 55.5% | 56.8% | 68.0% |
| **FewshotT5-PPDAC** | **78.4%** | **80.9%** | **86.3%** |
| FewshotT5-PPDAC(w/o. PF) | 61.5% | 66.4% | 75.5% |
| FewshotT5-PPDAC(w/o. PS) | 59.6% | 67.1% | 77.5% |
| FewshotT5-PPDAC(w/o. PF & PS) | 55.7% | 63.5% | 72.6% |
| FewshotT5 | 47.8% | 56.6% | 64.6% |

Table 4: The table presents the ablation study results on the SQuAD1.1 dataset for PPDAC in a few-shot learning context. The notation (w/o. PF) signifies the removal of the Paraphrase Filter from PPDAC, highlighting its impact. Similarly, (w/o. PS) indicates the exclusion of the Paraphrase Selector, allowing assessment of its individual contribution. Lastly, (w/o. PF & PS) refers to the scenario where both Paraphrase Filter and Paraphrase Selector are removed, providing insights into their combined effects on enhancing model performance.

CCL 2024

mentation and compared it to the baseline fine-tuning process without PPDAC. As evident from Table 3, while there is an observable increase in training time during the fine-tuning process, this increase is attributable to the data augmentation, with the enhanced data volume reaching 2 to 3 times that of the original dataset, and remains within an acceptable range. Notably, the model evaluation time is unaffected by the data augmentation component. The extra time needed for the data augmentation process primarily depends on the volume of the original data to be augmented. The additional space requirements for our module arise from the deployment of the paraphrase and filter models. The paraphrase model itself comprises 223 M parameters. In contrast, the adequacy filter and fluency filter are characterized by their parameter volumes of 109.5 M and 369 M, respectively. Overall, the increase in time and space requirements after applying the PPDAC component is marginal, reinforcing PPDAC's utility as a resource-efficient component in the EQA model training workflow.

### 4.6 Ablation Study

To elucidate the individual contributions of the Plug-and-Play Data Augmentation Component (PPDAC) to enhancing few-shot Extractive Question Answering (EQA) performance, we meticulously conducted an ablation study. This investigation, now encompassing both the paraphrase selector (PS) module and the Paraphrase Filter (PF) module, was performed across varying data scales—16, 64, and 128 shots within the SQuAD1.1 dataset. The paraphrase selector (PS) module and the Paraphrase Filter (PF) module are integral to enhancing the model's performance by ensuring the generated paraphrases not

only retain semantic fidelity to the original questions but also meet quality standards in terms of relevance and coherence. The PS module is crucial for selecting the most appropriate paraphrases by evaluating their alignment with the original question, whereas the PF module filters out paraphrases that do not meet predetermined quality criteria, thereby streamlining the data augmentation process for optimized few-shot learning outcomes. The findings, detailed in Table 4, are pivotal for understanding the nuanced impact of each component.

1) **Splinter-PPDAC(w/o. PS/PF)**: This variant employs the Splinter model integrated with PPDAC, excluding either the PS or PF module. This modification is designed to individually assess the impact of removing each module on the model's performance.

2) **FewshotBART-PPDAC(w/o. PS/PF)**: Similarly, this setup uses the FewshotBART framework with PPDAC, removing either the PS or PF module to evaluate their individual contributions to the model's architecture and effectiveness.

3) **FewshotT5-PPDAC(w/o. PS/PF)**: Following the same approach, this configuration employs the FewshotT5 framework with PPDAC, omitting either the PS or PF module to explore their separate roles in improving model performance.

From Table 3, it is evident that the performance of all models declines universally when any one components are removed, the simultaneous exclusion of both the Paraphrase Filter (PF) and Paraphrase Selector (PS) from PPDAC results in the most pronounced performance deterioration. For instance, within the FewshotT5-PPDAC model using 16 training examples, the removal of PF and PS modules led to performance drops of 16.9% and 18.8%, respectively, with their combined removal causing a 22.7% reduction. Despite these setbacks, the modified models continue to outperform their original configurations across all training sizes. Crucially, eliminating the PF module typically results in greater performance degradation than removing the PS, leading to the conclusion that the PF module's role in ensuring the quality of paraphrase generation is more critical than the PS component's contribution to selecting optimal paraphrases. This underscores the PF module's paramount importance in the enhancement of few-shot EQA performance through quality control over quantity.

Despite these performance decrements, models stripped of both PS and PF, relying solely on the parrot generator component, still surpassed baseline models. This underscores the inherent value of the PPDAC framework in improving EQA outcomes, even in its most reduced form. Among the components, the PS module emerges as particularly crucial, significantly impacting model performance by optimizing paraphrase selection for training efficiency and accuracy.

This detailed analysis not only highlights the indispensable contributions of the PS and PF modules to the PPDAC's efficacy but also sets a foundation for further exploration into refining EQA methodologies within few-shot learning paradigms.

## 5 Conclusion

In this work, we introduce a novel Plug-and-Play Data Augmentation Component (PPDAC) designed to enhance the training of pre-trained language models for question-answering (QA) tasks. By generating paraphrases ($q'$) of the original questions within (question, passage, answer) triples, PPDAC significantly enriches the training dataset, creating additional ($q'$, passage, answer) combinations. This innovative approach is particularly beneficial in few-shot learning scenarios, where data scarcity often hampers model performance. Our comprehensive experiments across various QA baselines demonstrate the efficacy of PPDAC, with notable improvements observed even in extremely data-constrained environments. Specifically, in a 16-shot learning setup on the SQuAD dataset, PPDAC facilitated performance enhancements of 3.6%, 1.9%, and 30.6% for the Splinter, FewshotBART, and FewshotT5 models, respectively. Remarkably, the integration of PPDAC with FewshotT5 not only achieved significant gains over baseline models but also surpassed existing state-of-the-art methods in few-shot settings by a substantial margin.

Furthermore, our detailed ablation studies shed light on the critical roles of PPDAC's components, particularly the Paraphrase Filter (PF) and Paraphrase Selector (PS). The studies reveal that removing any component leads to a universal decline in model performance, with the simultaneous exclusion of both PF and PS resulting in the most pronounced performance drops. However, even in the absence of

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1320-1333, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

these components, the models equipped with the remaining paraphrase generation capabilities of PPDAC continued to outperform their original configurations. This underscores the transformative impact of PPDAC on enhancing QA model performance in few-shot learning contexts, with the PF module, in particular, emerging as a key contributor by prioritizing quality over quantity in paraphrase generation.

Looking ahead, we acknowledge the effectiveness of large language models (LLMs) in data augmentation. However, the innovation of our approach lies in its novel framework and process. While our current implementation utilizes a specific paraphrase generation model, we recognize its adaptability. In future iterations, we plan to incorporate the latest LLMs for paraphrase generation, emphasizing our innovative approach tailored for question-answering tasks.

The Plug-and-Play Data Augmentation Component (PPDAC) currently enhances question paraphrasing within the (question, passage, answer) framework. Future research may expand this approach to include passage paraphrasing, thereby enriching the training dataset and establishing a more solid foundation for fine-tuning pre-trained language models. This expansion aims to advance few-shot learning in question answering, steering towards more sophisticated and effective natural language understanding systems.

## Acknowledgements

This work was supported in part by the NSFC (No. 62266023), the Science and Technology Research Project of the Department of Education of Jiangxi Province (GJJ210325, GJJ2200354). We would like to thank anonymous reviewers for the insightful comments. Mingwen Wang is the corresponding author.

## References

Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Jianfeng Gao, Songhao Piao, Ming Zhou, et al. 2020. Unilmv2: Pseudo-masked language models for unified language model pre-training. In *International conference on machine learning*, pages 642–652. PMLR.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Or Castel, Ori Ram, Avia Efrat, and Omer Levy. 2021. How optimal is greedy decoding for extractive question answering? *arXiv e-prints*, pages arXiv–2108.

Rakesh Chada and Pradeep Natarajan. 2021. Fewshotqa: A simple framework for few-shot learning of question answering tasks using pre-trained text-to-text models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6081–6090.

Prithiviraj Damodaran. 2021. Parrot: Paraphrase generation for nlu. *v1.0*.

Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. 2017. Searchqa: A new q&a dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*.

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. Mrqa 2019 shared task: Evaluating generalization in reading comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 1–13.

Michael Glass, Alfio Gliozzo, Rishav Chakravarti, Anthony Ferritto, Lin Pan, GP Shrivatsa Bhargav, Dinesh Garg, and Avirup Sil. 2020. Span selection pre-training for question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2773–2782.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, pages 5376–5384. Institute of Electrical and Electronics Engineers Inc.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

DP Kingma. 2014. Adam: a method for stochastic optimization. In *Int Conf Learn Represent*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Patrick Lewis, Ludovic Denoyer, and Sebastian Riedel. 2019. Unsupervised question answering by cloze translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4896–4910.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *International Conference on Learning Representations*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Ori Ram, Yuval Kirstain, Jonathan Berant, Amir Globerson, and Omer Levy. 2021. Few-shot question answering by pretraining span selection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3066–3079.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. Newsqa: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16(1):138.

J. Wang, Chengyu Wang, Minghui Qiu, Qiuhui Shi, Hongbin Wang, Jun Huang, and Ming Gao. 2022. Kecp: Knowledge enhanced contrastive prompting for few-shot extractive question answering. In *Conference on Empirical Methods in Natural Language Processing*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Ningyu Zhang, Shumin Deng, Xu Cheng, Xi Chen, Yichi Zhang, Wei Zhang, Huajun Chen, and Hangzhou Innovation Center. 2021. Drop redundant, shrink irrelevant: Selective knowledge injection for language pretraining. In *IJCAI*, pages 4007–4014.

| Model | SQuAD | TriviaQA | NQ | NewsQA | SearchQA | HotpotQA | BioASQ | TextbookQA |
|---|---|---|---|---|---|---|---|---|
| *16 Examples* | | | | | | | | |
| RoBERTa | 7.7±4.3 | 7.5±4.4 | 17.3±3.3 | 1.4±0.8 | 6.9±2.7 | 10.5±2.5 | 16.7±7.1 | 3.3±2.1 |
| RoBERTa-PPDAC | **10.1±1.1** | 5.7±1.4 | **19.1±2.8** | **5.9±0.6** | **8.1±1.8** | **11.0±2.0** | **19.5±1.0** | 3.2±1.5 |
| SpanBERT | 12.5±5.7 | 12.8±5.4 | 19.7±3.6 | 6.0±1.6 | 13.0±4.2 | 12.6±3.6 | 22.0±4.6 | 5.6±2.5 |
| SpanBERT-PPDAC | **20.2±6.9** | **16.3±2.1** | **20.7±3.7** | **11.9±0.4** | **22.1±1.6** | **13.4±0.8** | **23.7±0.6** | **7.8±1.2** |
| Splinter | 54.6±6.4 | 18.9±4.1 | 27.4±4.6 | 20.8±2.7 | 26.3±3.9 | 24.0±5.0 | 28.2±4.9 | 19.4±4.6 |
| Splinter-PPDAC | **58.2±2.5** | **21.4±4.8** | 26.9±2.2 | 19.3±2.2 | **28.7±2.1** | **28.8±2.6** | **28.8±1.0** | 19.4±3.9 |
| FewshotBART | 55.5±2.0 | 50.5±1.0 | 46.7±2.3 | 38.9±0.7 | 39.8±0.04 | 45.1±1.5 | 49.4±0.02 | 19.9±1.25 |
| FewshotBART-PPDAC | **57.4±0.8** | 47.1±0.1 | **47.7±0.3** | 37.8±0.6 | **45.9±1.2** | **60.4±1.1** | **49.9±1.8** | **20.1±1.0** |
| FewshotBARTL | 68.9±2.7 | 65.2±1.8 | 60.4±2.0 | 48.4±2.2 | 47.8±5.4 | 58.0±1.8 | 63.0±1.1 | 37.7±3.7 |
| FewshotBARTL-PPDAC | **73.2±1.5** | 59.6±3.2 | **60.5±2.4** | **49.2±0.8** | **59.5±1.0** | **59.3±2.4** | **64.6±0.8** | 37.4±1.3 |
| FewshotT5 | 47.8±6.9 | 50.6±4.9 | 28.5±14.5 | 26.8±2.7 | 37.0±3.3 | 44.9±3.5 | 46.3±5.9 | 25.9±5.0 |
| FewshotT5-PPDAC | **78.4±0.2** | **51.4±2.0** | **58.5±1.0** | **52.2±0.5** | **43.4±2.1** | **55.5±0.3** | **58.6±0.4** | **32.1±2.5** |
| KECP | 67.1±4.7 | 24.8±2.4 | 31.9±2.2 | 28.4±1.9 | 35.3±2.4 | 33.9±2.0 | - | - |
| *32 Examples* | | | | | | | | |
| RoBERTa | 18.2±5.1 | 10.5±1.8 | 22.9±0.7 | 3.2±1.7 | 13.5±1.8 | 10.4±1.9 | 23.3±6.6 | 4.3±0.9 |
| RoBERTa-PPDAC | **26.4±1.3** | **10.6±1.6** | **23.8±1.6** | **8.9±0.9** | **15.6±1.0** | **11.8±1.2** | **27.6±1.2** | **4.9±0.2** |
| SpanBERT | 19.0±4.6 | 19.0±4.8 | 23.5±0.9 | 7.5±1.3 | 20.1±3.9 | 14.4±2.9 | 32.5±3.5 | 7.4±1.1 |
| SpanBERT-PPDAC | **39.9±2.9** | **20.3±0.7** | **28.1±1.6** | **17.4±0.8** | **22.3±2.1** | **21.9±3.3** | **36.1±4.1** | **19.0±3.5** |
| Splinter | 59.2±2.1 | 28.9±3.1 | 33.6±2.4 | 27.5±3.2 | 34.8±1.8 | 34.7±3.9 | 36.5±3.2 | 27.6±4.3 |
| Splinter-PPDAC | **64.3±1.1** | **28.8±3.5** | **34.6±1.2** | **27.7±2.1** | **35.6±1.2** | **39.5±1.2** | **37.0±0.8** | **27.7±3.5** |
| FewshotBART | 56.8±2.1 | 52.5±0.7 | 50.1±1.1 | 40.4±1.5 | 41.8±0.02 | 47.9±1.4 | 52.3±0.02 | 22.7±2.3 |
| FewshotBART-PPDAC | **60.4±1.9** | 51.3±1.5 | 50.1±0.9 | **41.6±0.3** | **49.7±0.5** | **61.5±0.7** | **56.3±2.6** | **26.7±2.0** |
| FewshotBARTL | 72.3±1.0 | 65.1±1.2 | 61.5±1.7 | 51.7±1.7 | 58.3±1.5 | 60.4±0.2 | 67.8±1.0 | 37.7±9.8 |
| FewshotBARTL-PPDAC | **76.0±1.0** | 58.1±2.2 | **62.2±0.7** | **52.5±0.7** | **63.5±0.9** | **62.2±0.7** | **70±0.4** | **44.1±0.3** |
| FewshotT5 | 56.6±1.5 | 50.2±9.0 | 37.5±12.5 | 33.2±4.6 | 48.4±5.6 | 53.6±1.4 | 57.7±4.2 | 29.8±2.6 |
| FewshotT5-PPDAC | **80.9±0.4** | 50.2±1.1 | **60.1±0.1** | **55.6±0.8** | **48.5±1.3** | **57.9±0.5** | **63.8±0.1** | **39.6±3.2** |
| *64 Examples* | | | | | | | | |
| RoBERTa | 28.4±1.7 | 12.5±1.4 | 24.2±1.0 | 4.6±2.8 | 19.8±2.4 | 15.0±3.9 | 34.0±1.8 | 5.4±1.1 |
| RoBERTa-PPDAC | **36.3±1.4** | **15.6±0.8** | **26.4±1.1** | **17.3±0.3** | **20.2±1.2** | **18.7±1.2** | **37.5±1.7** | **6.7±0.5** |
| SpanBERT | 33.6±4.3 | 22.8±2.6 | 28.4±1.8 | 8.8±2.4 | 26.7±2.9 | 21.8±1,5 | 43.9±4.5 | 7.4±1.2 |
| SpanBERT-PPDAC | **49.8±1.6** | **23.0±1.2** | **33.7±1.0** | **26.3±2.6** | **28.6±1.5** | **35.1±1.7** | **50.5±1.6** | **25.9±2.3** |
| Splinter | 65.2±1.4 | 35.5±3.7 | 38.2±2.3 | 37.4±1.2 | 39.8±3.6 | 45.4±2.3 | 49.5±3.6 | 35.9±3.1 |
| Splinter-PPDAC | **66.6±0.4** | **36.1±2.1** | **38.4±0.3** | **37.8±1.4** | **41.2±1.3** | **47.6±1.7** | **50±0.7** | **36.1±0.8** |
| FewshotBART | 61.5±2.3 | 50.8±2.2 | 53.0±0.5 | 42.7±2.2 | 46.1±2.9 | 51.2±1.0 | 61.8±2.8 | 27.6±1.8 |
| FewshotBART-PPDAC | **67.8±0.6** | **53.5±0.6** | 51.3±0.7 | **44.1±0.3** | **55.2±1.3** | **64.4±0.3** | **63.8±1.0** | **34.3±0.4** |
| FewshotBARTL | 73.6±1.9 | 64.6±1.4 | 63.0±2.1 | 53.5±0.9 | 65.5±2.4 | 62.9±1.6 | 73.9±0.8 | 45.0±1.7 |
| FewshotBARTL-PPDAC | **78.6±0.6** | 55.6±0.6 | 63.3±0.8 | **53.7±0.6** | **67.6±0.7** | **63.3±0.8** | 72.0±1.6 | **46.2±0.9** |
| FewshotT5 | 57.2±5.6 | 52.4±5.9 | 48.6±2.1 | 40.2±4.1 | 54.4±3.0 | 56.3±2.9 | 63.8±2.5 | 32.1±2.7 |
| FewshotT5-PPDAC | **83.5±0.2** | 50.8±0.7 | **62.5±0.7** | **58.8±0.8** | **54.4±0.2** | **61.4±0.3** | **68.7±1.4** | **47.5±0.9** |
| *128 Examples* | | | | | | | | |
| RoBERTa | 43.0±7.1 | 19.1±2.9 | 30.1±1.9 | 16.7±3.8 | 27.8±2.5 | 27.3±3.9 | 46.1±1.4 | 8.2±1.1 |
| RoBERTa-PPDAC | **45.2±4.8** | **19.2±1.1** | **33.4±1.1** | **28.0±1.4** | **27.9±0.8** | **30.2±0.7** | **49.4±2.5** | **9.0±0.6** |
| SpanBERT | 48.5±7.3 | 24.2±2.1 | 32.2±3.2 | 17.4±3.1 | 34.3±1.1 | 35.1±4.2 | 55.3±3.8 | 9.4±3.0 |
| SpanBERT-PPDAC | **53.9±3.0** | 19.3±1.8 | **34.1±1.5** | **27.3±3.6** | **35.1±0.9** | **39.8±1.0** | **57.5±0.5** | **27.8±2.1** |
| Splinter | 72.7±1.0 | 44.7±3.9 | 46.3±0.8 | 43.5±1.3 | 47.2±3.5 | 54.7±1.4 | 63.2±4.1 | 42.6±2.5 |
| Splinter-PPDAC | **73.8±0.7** | **45.1±1.0** | **48.8±0.4** | **43.6±0.9** | **48.0±2.1** | **55.3±0.5** | **64.4±2.0** | 37.9±2.0 |
| FewshotBART | 68.0±0.3 | 50.1±1.8 | 53.9±0.9 | 47.9±1.2 | 58.1±1.4 | 54.8±0.8 | 68.5±1.0 | 29.7±2.4 |
| FewshotBART-PPDAC | **71.6±1.0** | **53.9±1.2** | **54.1±0.2** | **49.9±0.08** | **59.1±0.2** | **67.1±0.6** | **72.4±0.2** | **37.8±0.4** |
| FewshotBARTL | 79.4±1.5 | 65.8±0.9 | 64.3±1.3 | 57.0±0.9 | 67.7±1.0 | 75.1±1.5 | 75.0±1.5 | 48.4±2.7 |
| FewshotBARTL-PPDAC | **79.9±0.5** | 57.5±0.8 | **65.4±0.8** | **57.2±1.0** | **68.1±0.5** | 65.4±0.8 | **77.2±0.9** | **50.5±2.1** |
| FewshotT5 | 64.6±6.1 | 51.7±3.1 | 47.0±4.6 | 40.0±1.9 | 57.0±4.5 | 56.1±3.7 | 68.2±3.6 | 33.6±2.1 |
| FewshotT5-PPDAC | **86.3±0.04** | **52.8±0.3** | **64.3±0.1** | **60.9±0.4** | **60.8±0.4** | **64.4±0.9** | **73.9±0.9** | **51.5±0.8** |

Table 5: The table spans training examples from 16 to 128, where data highlighted in bold indicates performance improvements post the application of the PPDAC component.

## A Complete Experimental Data Tables

The appendix presents the complete set of experimental data tables, which, due to their extensive length, are divided into two subsets for 16, 32, 64, 128 and 256, 512, 1024,which are respectively Table 5 and Table 6. In these tables, improvements in model performance following the application of PPDAC are highlighted in bold font. The baseline KECP data is included for comparative reference as per the original publication, given the absence of open-source code from the authors. The term ”FewshotBARTL” denotes the large version of the BART model within the Fewshot framework. Given its significantly larger parameter volume compared to BART-base and T5-base, it is included in the tables solely for comparative reference.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1320-1333, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

| Model | SQuAD | TriviaQA | NQ | NewsQA | SearchQA | HotpotQA | BioASQ | TextbookQA |
|---|---|---|---|---|---|---|---|---|
| *256 Examples* | | | | | | | | |
| RoBERTa | 56.1±5.2 | 26.9±3.5 | 36.0±3.2 | 31.2±2.4 | 37.5±1.7 | 42.7±3.1 | 63.5±1.8 | 13.5±1.9 |
| RoBERTa-PPDAC | **56.2±2.8** | 26.9±2.1 | **36.1±1.1** | **32.0±1.4** | **37.9±0.8** | 42.2±0.7 | 63.4±2.5 | **14.3±0.6** |
| SpanBERT | 55.2±8.8 | 34.0±5.7 | 41.3±2.2 | 34.7±4.1 | 42.3±4.1 | 49.4±4.0 | 67.5±3.9 | 18.2±4.5 |
| SpanBERT-PPDAC | **55.9±3.0** | **34.1±1.8** | **42.1±1.5** | 34.3±3.6 | 42.1±0.9 | **49.8±1.0** | 67.5±0.5 | **31.6±2.1** |
| Splinter | 76.8±0.6 | 57.2±2.2 | 54.6±1.2 | 49.0±0.4 | 55.7±1.9 | 62.0±1.6 | 77.4±2.0 | 48.5±2.2 |
| Splinter-PPDAC | 76.8±0.5 | **57.3±1.0** | **54.8±0.4** | **49.6±0.9** | **56.0±2.1** | **63.3±0.5** | 77.4±2.0 | **48.9±2.0** |
| FewshotBART | 74.3±0.2 | 52.7±0.4 | 56.9±0.3 | 53.4±0.4 | 65.4±0.5 | 59.4±0.5 | 81.0±0.2 | 44.6±0.8 |
| FewshotBART-PPDAC | **74.6±0.2** | 52.7±0.1 | **57.9±0.2** | **53.8±0.4** | **64.7±0.1** | 59.4±0.5 | **82.1±0.7** | **47.2±0.9** |
| FewshotBARTL | 81.9±1.3 | 56.2±1.5 | 66.4±0.8 | 59.6±0.6 | 71.1±0.7 | 67.1±0.2 | 82.9±0.3 | 54.4±0.2 |
| FewshotBARTL-PPDAC | **82.2±0.04** | **61.6±0.6** | **66.6±0.1** | 59.6±0.3 | **71.4±0.4** | **67.3±0.1** | **83.6±1.2** | **57.2±0.3** |
| FewshotT5 | 87.4±0.3 | 55.4±0.7 | 65.6±0.1 | 61.8±0.1 | 65.9±0.4 | 66.5±0.4 | 77.9±0.1 | 50.3±1.5 |
| FewshotT5-PPDAC | **87.5±0.2** | **55.9±1.0** | **65.8±0.3** | **62.3±0.4** | **66.0±0.2** | 66.3±0.4 | **80.7±0.09** | **55.0±1.5** |
| *512 Examples* | | | | | | | | |
| RoBERTa | 67.3±0.7 | 38.7±3.8 | 46.7±2.2 | 41.5±2.2 | 46.9±1.6 | 56.7±1.3 | 77.0±1.9 | 27.0±2.2 |
| RoBERTa-PPDAC | 67.2±0.5 | 38.7±2.8 | 46.7±1.2 | 41.5±1.2 | 46.9±0.6 | 56.0±0,8 | **77.3±0.9** | **27.5±1.2** |
| SpanBERT | 70.0±4.3 | 44.2±2.9 | 51.5±1.8 | 42.4±2.6 | 53.9±3.2 | 61.6±1.7 | 80.3±3.0 | 33.7±3.4 |
| SpanBERT-PPDAC | **70.9±2.0** | 44.2±1.9 | 51.5±1.0 | 42.4±1.6 | 53.9±2.2 | **61.7±0.7** | **80.4±1.0** | **34.0±1.4** |
| Splinter | 80.1±0.4 | 61.9±1.8 | 61.4±1.1 | 53.2±0.9 | 63.1±1.6 | 66.2±0.6 | 84.8±0.9 | 54.2±1.7 |
| Splinter-PPDAC | **80.8±0.7** | 61.9±0.8 | 61.4±1.0 | 53.2±0.5 | 63.1±0.6 | **66.4±0.6** | **85.0±0.9** | **54.5±0.7** |
| FewshotBART | 77.1±0.2 | 51.9±0.6 | 59.3±0.1 | 56.4±0.3 | 67.3±0.2 | 63.5±0.7 | 87.6±0.8 | 50.4±0.04 |
| FewshotBART-PPDAC | **77.2±0.3** | **52.8±0.3** | **60.1±0.09** | **56.5±0.1** | **67.7±0.2** | 63.5±0.6 | 87.6±0.3 | **54.9±0.1** |
| FewshotBARTL | 83.3±0.8 | 60.4±1.4 | 66.4±0.8 | 60.8±0.09 | 73.0±0.4 | 70.3±0.4 | 84.4±0.1 | 55.7±1.2 |
| FewshotBARTL-PPDAC | 83.1±0.1 | **63.2±0.5** | **67.4±0.2** | **61.9±0.2** | 72.8±0.8 | 69.6±0.8 | **86.9±0.3** | **57.5±0.5** |
| FewshotT5 | 87.9±0.1 | 57.7±0.2 | 67.8±0.2 | 63.5±0.08 | 69.6±0.1 | 68.8±0.4 | 83.0±0.1 | 56.7±0.6 |
| FewshotT5-PPDAC | **88.5±0.1** | **58.1±0.4** | **68.0±0.4** | **63.9±0.1** | **69.8±0.1** | 68.8±0.4 | **84.1±0.9** | **58.8±0.1** |
| *1024 Examples* | | | | | | | | |
| RoBERTa | 73.8±0.8 | 46.8±0.9 | 54.2±1.1 | 47.5±1.1 | 54.3±1.2 | 61.8±1.3 | 84.1±1.1 | 35.8±2.0 |
| RoBERTa-PPDAC | 73.2±0.8 | **47.1±1.1** | **54.6±1.1** | 47.3±1.4 | 54.1±0.8 | **62.0±0.7** | **84.3±1.0** | **36.0±0.6** |
| SpanBERT | 77.8±0.9 | 50.3±4.0 | 57.5±0.9 | 49.3±2.0 | 60.1±2.2 | 67.4±1.6 | 89.3±0.6 | 42.3±1.9 |
| SpanBERT-PPDAC | **77.9±0.3** | **50.4±1.8** | **58.5±0.5** | **50.1±3.6** | **60.5±0.9** | 67.3±1.0 | 89.3±0.6 | 42.1±2.1 |
| Splinter | 82.8±0.8 | 64.8±0.9 | 65.5±0.5 | 57.3±0.8 | 67.3±1.3 | 70.3±0.8 | 91.0±1.0 | 54.5±1.5 |
| Splinter-PPDAC | 82.8±0.7 | **64.9±1.0** | **66.2±0.4** | 57.0±0.9 | 67.3±1.1 | **70.4±0.5** | 91.0±0.5 | **54.7±1.3** |
| FewshotBART | 79.6±0.09 | 55.5±0.7 | 61.4±0.1 | 58.9±0.6 | 70.0±0.04 | 65.8±0.7 | 92.2±0.8 | 50.0±0.2 |
| FewshotBART-PPDAC | **79.7±0.1** | 55.5±0.9 | **62.4±0.1** | 58.7±0.1 | **70.5±0.2** | 65.0±0.6 | **92.5±0.2** | **51.8±0.4** |
| FewshotBARTL | 83.0±0.2 | 62.1±0.1 | 68.2±0.2 | 62.2±0.4 | 73.4±0.8 | 66.5±0.1 | 91.0±0.3 | 52.5±0.5 |
| FewshotBARTL-PPDAC | **84.5±0.5** | **64.1±0.4** | **69.0±0.4** | **63.6±0.2** | **74.4±0.6** | **71.5±0.2** | **91.7±0.2** | **58.9±0.4** |
| FewshotT5 | 88.6±0.2 | 60.0±0.5 | 69.4±0.3 | 64.7±0.2 | 73.0±0.2 | 70.2±0.2 | 86.9±0.2 | 59.7±0.5 |
| FewshotT5-PPDAC | **89.0±0.09** | **61.0±0.08** | 69.4±0.08 | **64.9±0.1** | **73.1±0.1** | **71.2±0.2** | **87.8±0.1** | **60.1±0.6** |

Table 6: This table encompasses training examples ranging from 256 to 1024, with data shown in bold representing performance enhancements following the incorporation of the PPDAC component.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1320-1333, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

# Sentence-Space Metrics (SSM) for the Evaluation of Sentence Comprehension

**Jieyu Lin, Honghua Chen, Nai Ding***
Key Laboratory for Biomedical Engineering of Ministry of Education,
College of Biomedical Engineering and Instrument Sciences, Zhejiang University, Hangzhou, China
{jieyu_lin,honghuachen,ding_nai}@zju.edu.cn

## Abstract

It is a fundamental challenge to evaluate whether a model can truly capture the meaning of sentences. Evaluation of whether a model well captures the meaning of individual words, however, can be effectively achieved by analyzing whether the model encodes words in a vector space where semantically similar words form clusters. Inspired by this approach, we propose the Sentence-Space Metrics (SSM) to evaluate model interpretation of sentences, and the sentence space is constructed based on the pairwise entailment relationships between all sentence pairs within a sentence pool. We use three metrics to evaluate a sentence space, i.e., (1) sparsity, (2) clustering of related sentences, and (3) similarity with the sentence space measured from humans. The SSM is applied to evaluate 20 models, including ChatGPT, 18 BERT-family models fine-tuned for Natural Language Inference (NLI) task, as well as SimCSE, a sentence representation model. The SSM reveals dramatic differences among models: Although all models achieve high accuracy on standard NLI datasets such as MNLI, none of them mirrors the human behavior under the SSM. These results demonstrate that, compared with traditional accuracy measures, the SSM considers pairwise relationships between hundreds of sentences and therefore provide a more fine-grained evaluation of model interpretation of sentences.

## 1 Introduction

How to represent the meaning of a word or a sentence is a classic question in philosophy, linguistics, and psychology (Rumelhart, 1986; Lund and Kevin, 1997; Martin, 2007; Mikolov et al., 2013b; Mikolov et al., 2013a). Converging evidence from Natural Language Processing (NLP) studies and human neuroimaging studies suggests that the meaning of a word can be well described by its relationship with other words. For example, word embedding algorithms project words into a semantic space in which words with more similar meanings are more closely located (Başkaya et al., 2013; Baroni et al., 2014; Th et al., 2015; Wang et al., 2019). Similarly, human neuroimaging studies have shown that semantic space for words can be constructed based on human rating of the similarities between words and predicts how similarly the brain responds to different words (Wang et al., 2018; Wang et al., 2020). In human language, words are only the basic elements to express meaning, while the unique expressive power relies on the construction of sentences - We use sentences to represent compositional semantic structures, including events and propositions. The complexity and abundancy in sentence meaning renders the evaluation of sentence meaning much more challenging than that of word meaning (Muennighoff et al., 2023).

Here, inspired by the analysis of word semantic space, we evaluate how well an NLP model captures sentence meaning by constructing a sentence space. Each sentence is a point in the vectorial space, and the distance between two sentences represents how strongly the two sentences are related. Just like the relationship between two words can be measured based on different features, e.g., animacy, abstractness, and part of speech information, the relationship between two sentences can be measured in many different ways. Here, we use the graded entailment score to quantify the relationship between two

* Corresponding author: Nai Ding

©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1334-1350, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

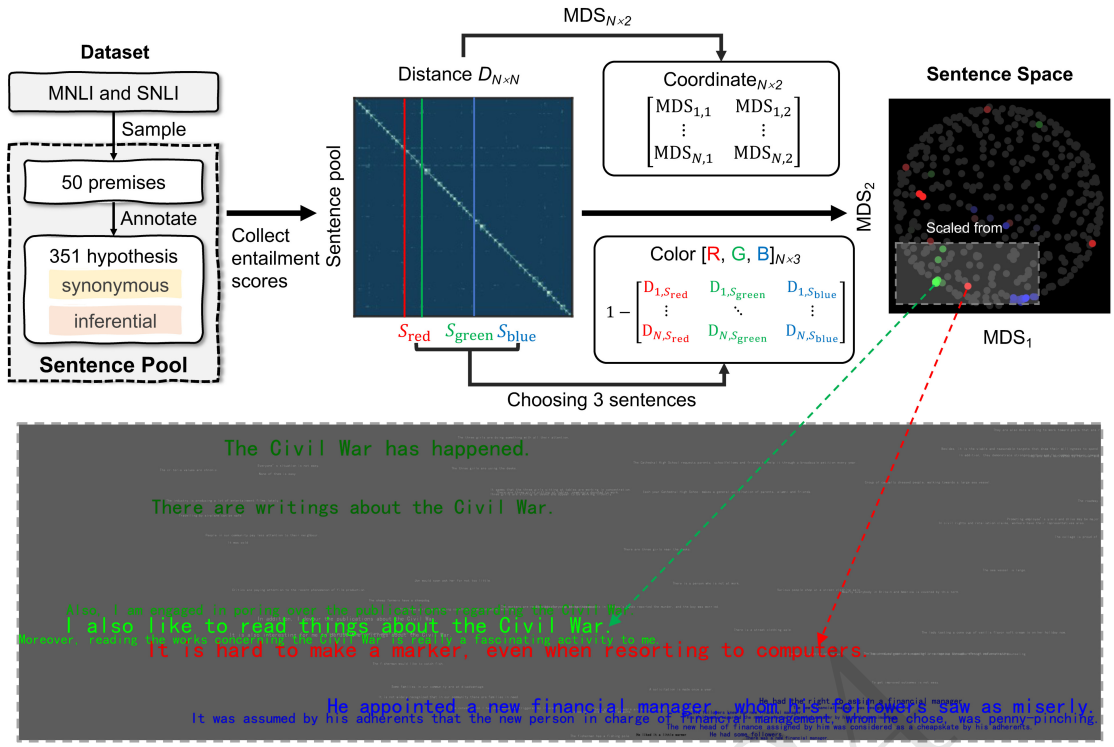



Figure 1: Construction of a semantic space for sentences. We construct a sentence pool, collect pairwise entailment scores to build a distance matrix, and use multidimensional scaling (MDS) to visualize the space. Each point in the sentence space in the upper right corner represents a sentence, and its color, i.e., RGB channels, illustrates how the sentence relates to 3 basis sentences that are randomly selected from the sentence pool. A portion of the sentence space is zoomed in in the lower panel, in which the font size is proportional to the shortest distance to a basis sentence – Larger fonts for sentences that are close to a basis sentence. For this illustration, the 3 basis sentences are: (R) Making a marker is a complicated task, even with modern computer assistance; (G) I also like to read things about the Civil War; (B) He appointed a new financial manager, whom his followers saw as miserly.

sentences (Jiang and de Marneffe, 2022). In other words, if a sentence could entail another sentence, the two sentences should have a shorter distance in the sentence space. We consider the semantic entailment score instead of other semantic similarity measures (Li et al., 2006; Mihalcea et al., 2006; Agirre et al., 2012) since semantic entailment is relatively more clearly defined while semantic similarity can be judged based on different sets of features (Deshpande et al., 2023). All the SSM, however, can be easily applied to sentence spaces constructed based on other measures.

Semantic entailment, also referred to as the Natural Language Inference (NLI) task, is widely used to evaluate the sentence comprehension ability of models (Bowman et al., 2015; Williams et al., 2018). Previous studies, however, mostly focus on the entailment relationship between pairs of sentences that are constructed either manually (Bowman et al., 2015; Williams et al., 2018; Nie et al., 2020) or based on templates (McCoy et al., 2019; Luo et al., 2022). For example, large datasets such as MNLI (Bowman et al., 2015) and SNLI (Williams et al., 2018) have been constructed by collecting premises from corpora, and asking annotators to construct hypotheses that have the required relationship with the premise (i.e., entailment, contradiction, or neutral). Transformers-based language models have achieved high accuracy on mainstream NLI datasets (Devlin et al., 2019; Liu et al., 2019; He et al., 2021), but their performance is sensitive to data corruption, distribution shift, and data manipulation (Poliak et al., 2018; Sanchez et al., 2018; Wallace et al., 2019a; Jin et al., 2020). Consequently, it is of significant interest to design more comprehensive methods to evaluate model performance on the NLI task (Naik et al., 2018; McCoy et al., 2019; Bartolo et al., 2020; Kiela et al., 2021; Liu et al., 2023). The SSM proposed here differs from

| Premise | Hypothesis |
|---|---|
| Making a marker is a complicated task, even with modern computer assistance. | Producing a marker is complex, even with the help of computers. (0.93, 0.83) |
| Making a marker is a complicated task, even with modern computer assistance. | Producing a marker is even trickier without computers. (0.80, 0.39) |
| A fisherman with his friend is setting up his pole. | A fisherman and his friend are preparing his fishing rod. (0.7, 0.76) |
| A fisherman with his friend is setting up his pole. | The fisherman would like to catch fish. (0.76, 0.33) |

Table 1: Examples of human-written hypotheses based on 2 premises. The numbers in parentheses represent the human entailment scores of the sentence pairs as well as the reversed pairs.

previous methods to evaluate sentence interpretation by jointly considering the relationships between all possible pairs of sentences within a sentence pool. Full paring of sentences naturally introduces sentence pairs that have diverse relationships, some of which can violate the heuristic cues in current large NLI datasets.

The procedure to construct a sentence space is illustrated in Figure 1. We constructed a sentence pool in which sentences fall into 50 semantically related clusters (6-11 sentences per cluster), calculated the distance between these sentences, i.e., the graded entailment score, and analyzed the space defined by the distance matrix. An example sentence space is visualized in Figure 1. We applied the method to test 20 language models, as well as human participants for comparison. For a model that well captures sentence meaning, we expect the following three properties. First, short distance should be sparse since most sentence pairs are unrelated. Second, semantically related sentences should form clusters. Third, the model sentence space should align with the human sentence space. Therefore, we use (1) sparsity, (2) clustering, and (3) similarity with human to evaluate a sentence space.

## 2 Method

### 2.1 Dataset

We construct a sentence pool, $S$, and perform the SSM based on $S$. The sentence pool $S$ differs from standard NLI datasets in three ways. First, since the SSM aims to examine the clustering of semantically related sentences in a sentence space, sentences in $S$ should form clusters, i.e., groups of sentences that are strongly related. Second, since sentence entailment is a directional measure, we distinguish *synonymous* sentences, i.e., two sentences can entail each other, and *inferential* sentences, i.e., one sentence can entail the other but not vice versa. Third, if there are $N$ sentences in $S$, the SSM considers all possible pairs from $S$, i.e., $N^2$ pairs of sentences. Therefore, $N$ must be relatively small to make the annotation of the entailment relationships between $N^2$ pairs feasible.

We initialize $S$ by randomly selecting 80 premises from the dev split of SNLI and MNLI under the constraints that the sentence contains 5-20 words. All premises are proofread to ensure that they are grammatical, meaningful, and semantically irrelevant with each other. We then expand $S$ with hand-written hypotheses. The 80 premises are divided into ten nonoverlapping sets, each containing 8 premises. Each set is given to an annotator and ten annotators are recruited in total. For each premise, the annotator has to write 2-5 *synonymous* sentences and 3-5 *inferential* sentences. See Table 1 for example. Since it may be challenging to write the hypotheses for some premises, the annotator is only required to finish the task for 5 out of the 8 premises. All the annotators are proficient in English and have passed CET-6, a standardized English test in China.

Altogether, 351 sentences are constructed after we filter out duplicated sentences and sentences that contain less than 3 content words. These 351 sentences, together with the 50 premises (39 from MNLI and 11 from SNLI), constitute the sentence pool $S$. Finally, $S$ contains 401 sentences that form 50 semantically related clusters around the 50 initial premises (6-11 sentences per cluster). The 401 sentences

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1334-1350, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

lead to $401 \times 401$ sentence pairs, and each pair is denoted as $(s_i, s_j)$ for $s_i, s_j \in S$, where $s_i$ is the premise and $s_j$ is the hypothesis. The sentence pairs fall into 5 categories: The first two categories are *inferential* pairs and *synonymous* pairs, which include the initial premise and the corresponding inferential and synonymous hypotheses written by human annotators. Another two categories are *inferential (reversed)* pairs and *synonymous (reversed)* pairs, which switch the premise and hypothesis in *inferential* and *synonymous* pairs. Sentence pairs not falling into the four categories are referred to as *irrelevant* pairs. The entailment relationship between a sentence pair is characterized by a graded measure between 0 and 1, i.e., $R(i, j)$ (Chen et al., 2020b). If $s_i$ can entail $s_j$ with high probability, $R(i, j)$ is near 1. In contrast, if the two sentences are unrelated or contradict each other, $R(i, j)$ is near 0. The pairwise entailment scores $R(i, j)$ form a matrix $R \in \mathbb{R}^{N \times N}$, where $N$ is the sentence number in $S$, i.e., 401.

### 2.2 Human Entailment Score

We collect human entailment scores, i.e., $R(i, j)$ for human. Since it is time-consuming to annotate the entailment relationships between all possible $401 \times 401$ sentence pairs, we only annotate a subset of sentence pairs. This subset includes all *Inferential*, *Synonymous*, *Inferential (reversed)* and *Synonymous (reversed)* pairs (702 pairs in total), as well as *Irrelevant* pairs that are judged to have an entailment score higher than 0.5 by at least 9 of the 18 BERT-family models. In total, 3,855 pairs of sentences are selected. The instruction for annotators is "For each pair of sentences, can the second sentence be inferred by the first sentence?". They choose from four options: (A) cannot be inferred, (B) can be inferred with low probability, (C) can be inferred with high probability, or (D) can be inferred. For the four options, $R(i, j)$ is scored as 0, 0.3, 0.7, and 1, respectively. Each sentence pair is scored by 8-11 annotators. Annotations by annotators with entailment accuracy below 0.6 are removed. Finally, $R(i, j)$ for human is averaged across annotators, where outliers beyond 3 standard deviation from the mean are removed. For unlabeled data, we assign $R(i, j)$ as 1 for identical sentence pairs (since a sentence always entails itself), and as 0 for the remaining *Irrelevant* pairs.

### 2.3 Model

#### 2.3.1 BERT-family NLI models

We use BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and DeBERTa-v3 (He et al., 2021) as pre-trained models, and consider both the base version and large version. The pre-trained models are provided by Huggingface (Wolf et al., 2019) and fine-tuned based on an NLI dataset, which could be SNLI, MNLI, or a combination of SNLI, MNLI, and ANLI (Nie et al., 2020). For concision, we use ANLI to refer the combination of SNLI, MNLI, and ANLI in subsequent sections. During fine-tuning, each model takes in a pair of sentences $(s_i, s_j)$, formats it as $[CLS, s_i, SEP, s_j, SEP]$, and encodes them together. The final embedding of CLS token is run through a linear layer to obtain three probability scores $(P^e_{i,j}, P^c_{i,j}, P^n_{i,j})$, each denoting the probability of the relationship between the sentence pair being entailment, contradiction, or neutral (Devlin et al., 2019; Falke et al., 2019). The training details are shown in Appendix A. Here, we only use $P^e_{i,j}$ as the entailment score, i.e., $R(i, j)$.

#### 2.3.2 SimCSE

SimCSE provides a universal vectorial sentence representation (Gao et al., 2021). The SimCSE model used in this work is based on RoBERTa-large (Liu et al., 2019), and is fine-tuned on SNLI and MNLI with a contrastive loss (Chen et al., 2020a). The model encodes each sentence as a vector, and the entailment score, i.e., $R(i, j)$, is defined as the cosine similarity between a pair of sentences. SimCSE is a model to evaluate semantic textual similarity, instead of an NLI model. Therefore, the model cannot possibly discriminate the directionality of the entailment relationship. We consider this model since it can possibly describe the synonymity of sentences.

#### 2.3.3 ChatGPT

ChatGPT is a large language model developed by OpenAI upon the InstructGPT (Ouyang et al., 2022), and we use the GPT-3.5-turbo API. Here, we directly ask ChatGPT to judge the entailment relationship between a sentence pair under a zero-shot setting, using the multiple-choice questions as same as in the

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1334-1350, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

human experiment. Each sentence pair is evaluated through an independent query to ChatGPT, to avoid the influence of previous samples. The prompts are shown in Appendix B.

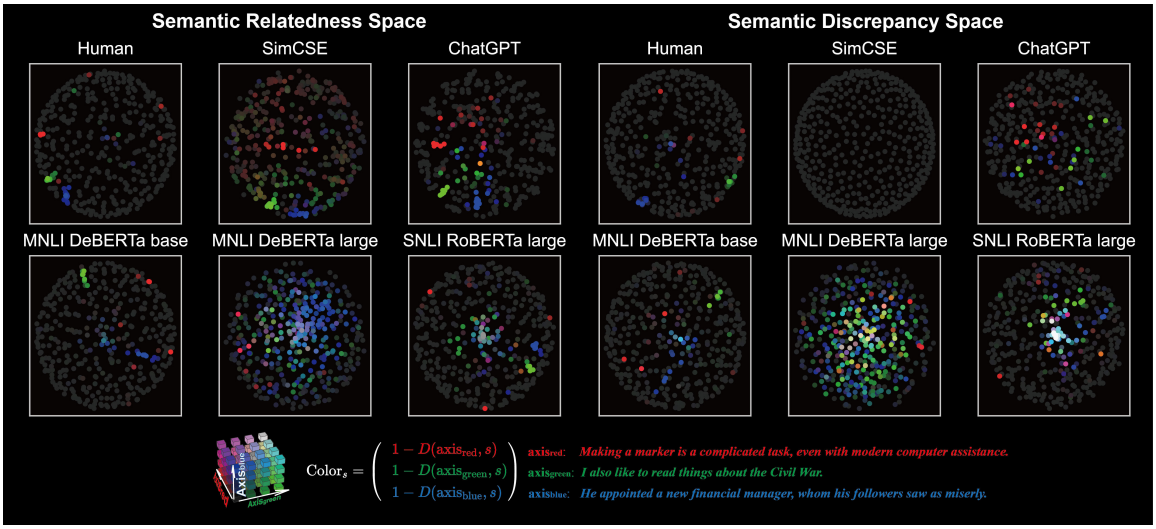



Figure 2: Visualization of the semantic relatedness space and semantic discrepancy space.

### 2.4 Entailment Accuracy

We turn the graded entailment score into a binary value to calculate the entailment accuracy. The BERT-family models are fine-tuned on the NLI task, and could predict a label, i.e., entailment or non-entailment (pooling the original 'contradiction' and 'neutral' labels). For human annotators, SimCSE, and ChatGPT, the relationship between a sentence pair is judged as entailment if and only if the entailment score is higher than 0.5. The mean human entailment score is taken as the ground truth.

### 2.5 Distance in Sentence Space

The entailment relationship between two sentences is directional – For a pair of sentences, one sentence may entail the other but not vice versa. In other words, there are two entailment scores $R(i,j)$ and $R(j,i)$ for each pair of sentences. We converted the two entailment scores into two unidirectional distances to construct two separate sentence spaces:

The semantic relatedness distance is defined as the average of the two entailment scores, i.e., $D_{rel}(i,j) = 1 - (R(i,j) + R(j,i))/2$, which describes general semantic similarity between two sentences. For example, the sentences on each row of Table 1 have relatively high entailment scores.

The semantic discrepancy distance is defined as the absolute difference of the two entailment scores, i.e., $D_{dis}(i,j) = 1 - |R(i,j) - R(j,i)|$, which describes how well the directional entailment relationship is captured. For example, on the even rows of Table 1, the premise entails the inferential sentence but not vice versa, resulting in a relatively large difference entailment score; while on the odd rows of Table 1, the sentences can entail each other and lead to a relatively low difference entailment score.

In the following, the calculations are applied to both $D_{rel}$ and $D_{dis}$, and the two distance matrices are also both referred to as $D$.

### 2.6 2D Visualization of Sentence Space

We visualize the sentence space by showing each sentence as a colored dot in a 2-D space. The coordinate of a sentence is determined by the 2-D multidimensional scaling (MDS) (Mead, 1992) based on $D$. The RGB color indicates its distance to each of 3 basis sentences that are randomly chosen from $S$. For example, if $s_{\text{green}}$ is selected as the basis for the green color, the value for the G channel is $1 - D(s_{\text{green}}, s_i)$ for $s_i \in S$. Consequently, sentences that have shorter distance to $s_{\text{green}}$ and much longer distance to $s_{\text{red}}$ and $s_{\text{blue}}$ will have greenish color on the graph. If a sentence is close to none of the three basis sentences, its color is black. In contrast, if the sentence is close to all three basis sentences, its color is white.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1334-1350, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

| Model | Dataset | Synonymous | Synonymous (reversed) | Inferential | Inferential (reversed) | Irrelevant |
|---|---|---|---|---|---|---|
| BERT base | ANLI | 0.748 | 0.771 | 0.650 | 0.791 | 0.460 |
| | MNLI | 0.733 | 0.756 | 0.645 | 0.791 | 0.720 |
| | SNLI | 0.664 | 0.763 | 0.686 | 0.786 | 0.382 |
| BERT large | ANLI | 0.840 | 0.847 | 0.714 | 0.809 | 0.417 |
| | MNLI | 0.794 | 0.832 | 0.664 | 0.795 | 0.691 |
| | SNLI | 0.802 | 0.832 | 0.714 | 0.768 | 0.388 |
| DeBERTa base | ANLI | 0.847 | 0.832 | 0.727 | 0.809 | 0.405 |
| | MNLI | 0.840 | 0.840 | 0.723 | 0.809 | 0.716 |
| | SNLI | 0.847 | 0.855 | 0.705 | 0.800 | 0.398 |
| DeBERTa large | ANLI | 0.840 | 0.840 | 0.723 | 0.809 | 0.436 |
| | MNLI | 0.809 | 0.855 | 0.686 | 0.800 | 0.692 |
| | SNLI | 0.855 | 0.847 | 0.718 | 0.805 | 0.419 |
| RoBERTa base | ANLI | 0.817 | 0.855 | 0.673 | 0.809 | 0.457 |
| | MNLI | 0.817 | 0.840 | 0.691 | 0.805 | 0.645 |
| | SNLI | 0.794 | 0.748 | 0.705 | 0.773 | 0.443 |
| RoBERTa large | ANLI | 0.802 | 0.824 | 0.718 | 0.809 | 0.530 |
| | MNLI | 0.824 | 0.817 | 0.691 | 0.805 | 0.495 |
| | SNLI | 0.878 | 0.832 | 0.700 | 0.827 | 0.414 |
| SimCSE | | 0.893 | 0.885 | 0.686 | 0.468 | 0.806 |
| ChatGPT | | 0.916 | 0.885 | 0.759 | 0.282 | 0.819 |
| Human | | 0.805 | 0.809 | 0.780 | 0.792 | 0.841 |

Table 2: Accuracy for the entailment judgement.

The three basis sentences are only used for visualization and the quantitative analyses are based on all sentences as detailed in the following.

### 2.7 Sentence-Space Metrics (SSM)

#### 2.7.1 Sparsity

In $S$, most sentence pairs are unrelated. Therefore, we expect a well-constructed sentence space to be sparse in the sense that most sentence pairs are unrelated and the distance between them should be near its maximal value, i.e., 1. We use the averaged $l^1$ norm (Hurley and Rickard, 2009) to quantify the sparsity of distance matrices $D \in \mathbb{R}^{N\times N}$. Formally, we calculate $\frac{\|1-D\|_1}{N^2}$. Smaller value indicates higher sparsity – The distance between most sentences is near 1.

#### 2.7.2 Clustering

Clustering of semantically similar words is often used as a method to evaluate word vector space (Baroni et al., 2014). Here, since the hypotheses are written based on 50 premises, they should fall into 50 clusters, with all the hypotheses written for the same premise and premise itself belonging to the same cluster. To capture how well a model clusters semantically related sentences, we calculate the Calinski-Harabasz Index (CH) (Caliński and JA, 1974). Specifically, for each sentence $s_i$ in $S$, its vector representation $d_i = [d_{i,1}, d_{i,2}, \cdots, d_{i,N}]$ is the corresponding row in the distance matrix, and it is classified into one of the 50 predefined clusters. The CH Index measures the ratio of the between-cluster isolation, the sum of the squared distance between the center of each cluster and the center of $S$, to the within-cluster coherence, the sum of the squared distance between each sentence and the center of its cluster. A larger value of CH indicates better cluster performance.

#### 2.7.3 Similarity with Human

We compare the similarity of sentence space between the model and human using the representational similarity analysis (RSA) score (Kriegeskorte et al., 2006; Kriegeskorte et al., 2008). The RSA concerns whether the distance matrix $D$ rated by human and models are consistent or not. Specifically, the RSA method flattens the upper diagonal part of $D$ into a vector and calculate the Pearson correlation coefficient (Pearson, 1896; Huitson et al., 1976; Rodgers et al., 1988) between the vectors.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1334-1350, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

| Model | Dataset | Sparsity (L1) | | Clustering (CH) | | Similarity (RSA) | |
|---|---|---|---|---|---|---|---|
| | | rel | dis | rel | dis | rel | dis |
| BERT base | ANLI | 0.04 | 0.06 | 6.09 | 3.15 | 0.63 | 0.55 |
| | MNLI | 0.02 | 0.02 | 7.69 | 4.10 | 0.76 | 0.67 |
| | SNLI | 0.10 | 0.13 | 5.87 | 2.99 | 0.44 | 0.40 |
| BERT large | ANLI | 0.06 | 0.09 | 5.66 | 2.95 | 0.61 | 0.54 |
| | MNLI | 0.04 | 0.06 | 6.12 | 3.94 | 0.52 | 0.39 |
| | SNLI | 0.09 | 0.13 | 5.60 | 2.93 | 0.46 | 0.39 |
| DeBERTa base | ANLI | 0.05 | 0.07 | 5.79 | 2.85 | 0.63 | 0.56 |
| | MNLI | 0.03 | 0.04 | 8.48 | 4.08 | 0.79 | 0.71 |
| | SNLI | 0.08 | 0.10 | 4.95 | 2.68 | 0.51 | 0.45 |
| DeBERTa large | ANLI | 0.06 | 0.09 | 5.47 | 2.92 | 0.57 | 0.51 |
| | MNLI | 0.13 | 0.15 | 6.34 | 3.92 | 0.26 | 0.29 |
| | SNLI | 0.06 | 0.08 | 6.29 | 2.85 | 0.59 | 0.49 |
| RoBERTa base | ANLI | 0.04 | 0.07 | 6.50 | 3.07 | 0.67 | 0.58 |
| | MNLI | 0.04 | 0.05 | 7.60 | 3.78 | 0.73 | 0.63 |
| | SNLI | 0.09 | 0.14 | 5.56 | 3.41 | 0.41 | 0.32 |
| RoBERTa large | ANLI | 0.04 | 0.06 | 7.74 | 3.31 | 0.73 | 0.64 |
| | MNLI | 0.06 | 0.09 | 5.40 | 2.87 | 0.58 | 0.50 |
| | SNLI | 0.08 | 0.13 | 4.48 | 2.48 | 0.46 | 0.37 |
| SimCSE | | 0.07 | / | 28.04 | / | 0.56 | / |
| ChatGPT | | 0.07 | 0.07 | 12.50 | 3.25 | 0.50 | 0.12 |
| Human | | 0.01 | 0.01 | 10.73 | 4.85 | 1.00 | 1.00 |

Table 3: Model performance under the SSM. Rel and dis represent the semantic relatedness space and the semantic discrepancy space, respectively.

## 3 Results

### 3.1 Entailment Accuracy

We first evaluate how accurately models could judge the entailment relationship between sentence pairs in our dataset (Table 2). Most models reach relatively high accuracy on human-written sentence pairs as well as the reversed pairs (*synonymous*, *synonymous (reversed)*, *inferential*, and *inferential (reversed)* pairs). For *synonymous* and *synonymous (reversed)* pairs, ChatGPT and SimCSE outperform BERT-family models but the difference is less than 15%. For *inferential (reversed)* pairs, however, ChatGPT and SimCSE perform much worse than BERT-family models (>30% difference). For SimCSE, $R(i,j) = R(j,i)$ and therefore the model cannot correctly judge the entailment relationship in *inferential (reversed)* pairs. ChatGPT, however, perform even worse than SimCSE in *inferential (reversed)* pairs, indicating a failure to detect bidirectional entailment relationship between sentences (see, e.g., the even rows of Table 1).

For *irrelevant* sentence pairs, the performance of BERT-family models is generally low, i.e., below 50% for more than half of the models, and their performance is not clearly higher for larger models (i.e., large vs. base model) and larger fine-tuning datasets (i.e., ANLI vs. MNLI/SNLI). The performance of SimCSE and ChatGPT is much higher than BERT-family models and is comparable to human performance.

### 3.2 Sentence-Space Metrics (SSM)

Next, we evaluate the models based on the SSM. We construct and visualize the sentence space following the procedure in Figure 1. Figure 2 separately visualize the semantic relatedness space and semantic discrepancy space for humans and a few representative models. The sentence space for other models is shown in Appendix C.

Since the three basis sentences are independently chosen and semantically irrelevant, in general, a sentence can possibly relate to at most one of them. In other words, we expect the color of most sentences to be black, and a few sentences could be red, green, or blue. For some BERT-family models, however, the sentence spaces tend to have numerous colored dots (Figure 2). More importantly, some dots are

CCL 2024

nearly white, indicating the implausible condition that these sentences have short distances to all three basis sentences. The visualization in Figure 2 is colored based on three randomly chosen basis sentences. To more comprehensively evaluate the sentence space properties, we evaluate the entire sentence space under the SSM.

#### 3.2.1 Sparsity

In the sentence pool, most sentence pairs are unrelated, and the distance between them is supposed to near its maximal value, i.e., 1. To characterize whether the distance is truly around 1 for most sentence pairs, we calculate the averaged $l^1$ norm of 1 minus the distance matrix (Table 3). As expected, the sentence space for humans has an averaged $l^1$ norm near 0, suggesting a highly sparse space. For models such as the MNLI BERT base, and MNLI DeBERTa base, the sentence space also has an averaged $l^1$ norm near 0, lower than the $l^1$ norm of SimCSE and ChatGPT. However, for some other models, e.g., MNLI DeBERTa large, the sentence space has a larger averaged $l^1$ norm, indicating a bias to judge unrelated sentences to have an entailment relationship.

#### 3.2.2 Clustering

In our sentence pool, 5-10 hypotheses are composed for each of the 50 premises. Hypotheses associated with the same premise are semantically related and should cluster in the sentence space. We quantify whether these sentences truly cluster using the CH Index. Specifically, we define 50 categories based on the 50 premises and separately calculate the CH Index in the semantic relatedness and discrepancy spaces (Table 3). A CH Index near 0 indicates that the 50 clusters are not separated, suggesting semantically related sentences are not more close to each other in the sentence space compared with irrelevant sentences. In contrast, a higher CH Index indicates better separability of the 50 clusters. In the semantic relatedness space, the CH Index is the highest for SimCSE, ChatGPT, and humans. In the semantic discrepancy space, the CH Index is the highest for humans and a couple of BERT-family models.

#### 3.2.3 Similarity with Human

Finally, we calculate the similarity of sentence space between the model and human by calculating the Pearson correlation coefficient between two distance matrices (Table 3). The results indicate that MNLI DeBERTa base is most similar to humans, followed by MNLI BERT base. The correlation between MNLI DeBERTa base and human reaches about 0.7 for both the semantic similarity and semantic discrepancy spaces. In semantic relatedness space, MNLI DeBERTa large perform the worst, suggesting that it could not well capture the degree of entailment. In semantic discrepancy space, ChatGPT perform the worst, suggesting that it cannot well capture the directional entailment relationship.

## 4 Related Work

### 4.1 Natural Language Inference

We consider a graded instead of categorical label, since Pavlick and Kwiatkowski (2019) and Zhang et al. (2021) found inherent disagreements in human textual inferences, and graded entailment score can capture more fine-grained relationship between sentences. Current NLI models are susceptible to distribution shifts and data manipulation (Hendrycks and Gimpel, 2017; Gururangan et al., 2018; Ebrahimi et al., 2018; Wallace et al., 2019a; Miller et al., 2020). To evaluate the models, previous studies have constructed challenging sets through, e.g., altering sentence structure (e.g., by changing the voice, clause structure, and negation etc.), replacing words with semantically similar or dissimilar words, altering sentence length by appending a tautology sentence (Naik et al., 2018; Glockner et al., 2018; McCoy et al., 2019; Luo et al., 2022; Liu et al., 2023). Previous studies have also used adversarial attack to test the robustness of models (Jia and Liang, 2017; Wallace et al., 2019a; Behjati et al., 2019; Jin et al., 2020). Consistent with our result, previous studies have also found that models tend to judge unrelated sentences to be entailment (Belinkov et al., 2019; Song et al., 2020; Lin et al., 2021; Luo et al., 2022).

To avoid predefined templates or fixed sentence patterns, other studies suggest to involve humans to construct challenging examples (Wallace et al., 2019b; Nie et al., 2020; Bartolo et al., 2020; Kiela et al., 2021). Our method only asks the annotators to generate a few hypotheses for each premise, not requiring

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1334-1350, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

the hypotheses to be challenging and therefore reducing the demand of human labor. The resulting sentence pairs exhibit natural variations in terms of syntactic, lexical, and semantic relationship. Since the SSM considers the relationship between all sentence pairs from a sentence pool, it involves more pairs of comparison. The large number of comparisons can reveal more detailed model performance, but also imposes a high labor cost for the labeling of ground truth. The cost of providing labels, however, is much lower than the cost to construct challenging samples.

Our results suggest that ChatGPT exhibits high accuracy on human-written pairs, but shows limited performance under the SSM. Qin et al. (2023) show that ChatGPT performs well but still underperforms fine-tuned models in two NLI datasets (i.e., RTE and CommitmentBank). Zhong et al. (2023) evaluate ChatGPT on GLUE benchmark, and find that ChatGPT have comparable performance with BERT-family models on NLI task. Some studies also suggest that ChatGPT has blind spots with certain types of samples (Basmov et al., 2023).

### 4.2 Sentence Vector Space

There are many ways to construct a sentence embedding. For example, some studies build upon the distributional hypothesis and learn a sentence embedding by predicting surrounding sentences (Kiros et al., 2015; Hill et al., 2016; Logeswaran and Lee, 2018). Other studies focus on deriving a sentence embedding by combining word embeddings (Mitchell and Lapata, 2008; Kalchbrenner et al., 2014; Wieting et al., 2016). Recently, contrastive training has been introduced to construct sentence embeddings (Gao et al., 2021; Ni et al., 2022). Previous studies evaluated the sentence embedding by a downstream task, e.g., semantic textual similarity. Here, however, we aim to utilize the sentence space to evaluate a downstream model.

## 5 Summary

We construct a sentence space to evaluate the sentence interpretation ability of models. The underlying philosophy of the method is that the meaning of a sentence is well described by its relationship with other sentences. Here, we use the graded entailment relationship between sentences as a distance measure to construct the sentence space, but the method can be readily used for other sentence similarity or relatedness measures. The evaluation results show that current models, including ChatGPT, still do not resemble the human ability to make language-based inference, even when the testing datasets do not involve challenging world knowledge or multi-media information. In sum, by constructing a sentence space and calculating metrics based on the space, we propose a new method that can more comprehensively evaluate sentence processing abilities of language models.

## 6 Limitations

Sentence spaces are only used for model evaluation, and future work will be directed toward explaining why model perform suboptimally in the sentence space and designing methods to improve model performance in the sentence space.

## Acknowledgements

The study is supported by the National Key Research and Development Program of China (No. 2021ZD0204105).

## References

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. In ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of*

*the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247, Baltimore, Maryland, June. Association for Computational Linguistics.

Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. 2020. Beat the AI: investigating adversarial human annotation for reading comprehension. *Trans. Assoc. Comput. Linguistics*, 8:662–678.

Osman Başkaya, Enis Sert, Volkan Cirik, and Deniz Yuret. 2013. AI-KU: Using substitute vectors and co-occurrence modeling for word sense induction and disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 300–306, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Victoria Basmov, Yoav Goldberg, and Reut Tsarfaty. 2023. ChatGPT and Simple Linguistic Inferences: Blind Spots and Blinds. *arXiv e-prints*, page arXiv:2305.14785, May.

Melika Behjati, SeyedMohsen MoosaviDezfooli, Mahdieh Soleymani Baghshah, and Pascal Frossard. 2019. Universal adversarial attacks on text classifiers. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*, pages 7345–7349. IEEE.

Yonatan Belinkov, Adam Poliak, Stuart M. Shieber, Benjamin Van Durme, and Alexander M. Rush. 2019. On adversarial removal of hypothesis-only bias in natural language inference. In Rada Mihalcea, Ekaterina Shutova, Lun-Wei Ku, Kilian Evang, and Soujanya Poria, editors, *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics, *SEM@NAACL-HLT 2019, Minneapolis, MN, USA, June 6-7, 2019*, pages 256–262. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, September. Association for Computational Linguistics.

Tadeusz Caliński and Harabasz JA. 1974. A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods*, 3:1–27, 01.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020a. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 13–18 Jul.

Tongfei Chen, Zhengping Jiang, Adam Poliak, Keisuke Sakaguchi, and Benjamin Van Durme. 2020b. Uncertain natural language inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8772–8779, Online, July. Association for Computational Linguistics.

Ameet Deshpande, Carlos E. Jimenez, Howard Chen, Vishvak Murahari, Victoria Graf, Tanmay Rajpurohit, Ashwin Kalyan, Danqi Chen, and Karthik Narasimhan. 2023. CSTS: conditional semantic textual similarity. *CoRR*, abs/2305.15093.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. Hotflip: White-box adversarial examples for text classification. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 31–36. Association for Computational Linguistics.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2214–2220. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6894–6910. Association for Computational Linguistics.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI systems with sentences that require simple lexical inferences. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 650–655. Association for Computational Linguistics.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 107–112. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: decoding-enhanced bert with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Dan Hendrycks and Kevin Gimpel. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning distributed representations of sentences from unlabelled data. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1367–1377, San Diego, California, June. Association for Computational Linguistics.

Alan Huitson, Olive Jean Dunn, and Virginia A. Clark. 1976. Applied statistics: Analysis of variance and regression. *The Statistician*, 25:236–236.

Niall P. Hurley and Scott T. Rickard. 2009. Comparing measures of sparsity. *IEEE Trans. Inf. Theory*, 55(10):4723–4741.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2021–2031. Association for Computational Linguistics.

Nanjiang Jiang and Marie-Catherine de Marneffe. 2022. Investigating reasons for disagreement in natural language inference. *Trans. Assoc. Comput. Linguistics*, 10:1357–1374.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is BERT really robust? A strong baseline for natural language attack on text classification and entailment. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8018–8025. AAAI Press.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pages 655–665. The Association for Computer Linguistics.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 4110–4124. Association for Computational Linguistics.

CCL 2024

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 3294–3302.

Nikolaus Kriegeskorte, Rainer Goebel, and Peter Bandettini. 2006. Information-based functional brain mapping. *Proceedings of the National Academy of Sciences*, 103(10):3863–3868.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. 2008. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, page 4.

Yuhua Li, David McLean, Zuhair Bandar, James O'Shea, and Keeley A. Crockett. 2006. Sentence similarity based on semantic nets and corpus statistics. *IEEE Trans. Knowl. Data Eng.*, 18(8):1138–1150.

Jieyu Lin, Jiajie Zou, and Nai Ding. 2021. Using adversarial attacks to reveal the statistical bias in machine reading comprehension models. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 2: Short Papers), Virtual Event, August 1-6, 2021*, pages 333–342. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Wei Liu, Ming Xiang, and Nai Ding. 2023. Adjective scale probe: Can language models encode formal semantics information? *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):13282–13290, Jun.

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Curt Burgess Lund and Kevin. 1997. Modelling parsing constraints with high-dimensional context space. *Language and cognitive processes*, 12(2-3):177–210.

Cheng Luo, Wei Liu, Jieyu Lin, Jiajie Zou, Ming Xiang, and Nai Ding. 2022. Simple but challenging: Natural language inference models fail on simple sentences. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 3449–3462. Association for Computational Linguistics.

Alex Martin. 2007. The representation of object concepts in the brain. *Annu. Rev. Psychol.*, 58:25–45.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3428–3448. Association for Computational Linguistics.

Andrew W. Mead. 1992. Review of the development of multidimensional scaling methods. *The Statistician*, 41:27–39.

Rada Mihalcea, Courtney D. Corley, and Carlo Strapparava. 2006. Corpus-based and knowledge-based measures of text semantic similarity. In *Proceedings, The Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference, July 16-20, 2006, Boston, Massachusetts, USA*, pages 775–780. AAAI Press.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun, editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 3111–3119.

John Miller, Karl Krauth, Benjamin Recht, and Ludwig Schmidt. 2020. The effect of natural distribution shift on question answering models. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 6905–6916. PMLR.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In Kathleen R. McKeown, Johanna D. Moore, Simone Teufel, James Allan, and Sadaoki Furui, editors, *ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, June 15-20, 2008, Columbus, Ohio, USA*, pages 236–244. The Association for Computer Linguistics.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. MTEB: massive text embedding benchmark. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 2006–2029. Association for Computational Linguistics.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Jianmo Ni, Gustavo Hernández Ábrego, Noah Constant, Ji Ma, Keith B. Hall, Daniel Cer, and Yinfei Yang. 2022. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1864–1874. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4885–4901. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Trans. Assoc. Comput. Linguistics*, 7:677–694.

Karl Pearson. 1896. Mathematical contributions to the theory of evolution. iii. regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society A*, 187:253–318.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In Malvina Nissim, Jonathan Berant, and Alessandro Lenci, editors, *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, *SEM@NAACL-HLT 2018, New Orleans, Louisiana, USA, June 5-6, 2018*, pages 180–191. Association for Computational Linguistics.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *CoRR*, abs/2302.06476.

Joseph Lee Rodgers, W Alan Nicewander, and David C. Blouin. 1988. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42:59–66.

David E Rumelhart. 1986. Learning internal representations by back-propagating errors. *Parallel distributed processing: Explorations in the microstructure of cognition*, pages 318–362.

Ivan Sanchez, Jeff Mitchell, and Sebastian Riedel. 2018. Behavior analysis of NLI models: Uncovering the influence of three factors on robustness. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1975–1985, New Orleans, Louisiana, June. Association for Computational Linguistics.

Congzheng Song, Alexander M. Rush, and Vitaly Shmatikov. 2020. Adversarial semantic collisions. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 4198–4210. Association for Computational Linguistics.

CCL 2024

Muneeb Th, Sunil Kumar Sahu, and Ashish Anand. 2015. Evaluating distributed word representations for capturing semantics of biomedical concepts. In Kevin Bretonnel Cohen, Dina Demner-Fushman, Sophia Ananiadou, and Jun'ichi Tsujii, editors, *Proceedings of the Workshop on Biomedical Natural Language Processing, BioNLP@IJCNLP 2015, Beijing, China, July 30, 2015*, pages 158–163. Association for Computational Linguistics.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019a. Universal adversarial triggers for attacking and analyzing NLP. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2153–2162. Association for Computational Linguistics.

Eric Wallace, Pedro Rodriguez, Shi Feng, Ikuya Yamada, and Jordan L. Boyd-Graber. 2019b. Trick me if you can: Human-in-the-loop generation of adversarial question answering examples. *Trans. Assoc. Comput. Linguistics*, 7:387–401.

Xiaosha Wang, Wei Wu, Zhenhua Ling, Yangwen Xu, Yuxing Fang, Xiaoying Wang, Jeffrey R Binder, Weiwei Men, Jia-Hong Gao, and Yanchao Bi. 2018. Organizational principles of abstract words in the human brain. *Cerebral Cortex*, 28(12):4305–4318.

Bin Wang, Angela Wang, Fenxiao Chen, Yuncheng Wang, and C.-C. Jay Kuo. 2019. Evaluating word embedding models: Methods and experimental results. *CoRR*, abs/1901.09785.

Xiaoying Wang, Weiwei Men, Jiahong Gao, Alfonso Caramazza, and Yanchao Bi. 2020. Two forms of knowledge representations in the human brain. *Neuron*, 107(2):383–393.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2016. Towards universal paraphrastic sentence embeddings. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Shujian Zhang, Chengyue Gong, and Eunsol Choi. 2021. Learning with different amounts of annotation: From zero to many labels. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7620–7632, Online and Punta Cana, Dominican Republic, nov. Association for Computational Linguistics.

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2023. Can chatgpt understand too? A comparative study on chatgpt and fine-tuned BERT. *CoRR*, abs/2302.10198.

CCL 2024

## Appendix A Training Details

| Parameter | Training Data | BERT | | RoBERTa | | DeBERTa | |
|---|---|---|---|---|---|---|---|
| | | base | large | base | large | base | large |
| Learning rate | MNLI | 2E-05 | 2E-05 | 2E-05 | 6E-06 | 2E-05 | 6E-06 |
| | SNLI | 3E-05 | 3E+05 | 2E-05 | 6E-06 | 2E-05 | 5E-06 |
| | ANLI | 3E-05 | 3E-05 | 2E-05 | 6E-06 | 2E-05 | 5E-06 |
| Train epochs | MNLI | 3 | 3 | 3 | 2 | 3 | 2 |
| | SNLI | 2 | 2 | 3 | 2 | 2 | 2 |
| | ANLI | 2 | 2 | 3 | 2 | 2 | 2 |
| Batch size | MNLI | 32 | 32 | 32 | 64 | 64 | 32 |
| | SNLI | 32 | 32 | 32 | 64 | 64 | 32 |
| | ANLI | 32 | 32 | 32 | 64 | 64 | 32 |
| Weight decay | MNLI | 0.01 | 0.01 | 0.10 | 0.00 | 0.00 | 0.00 |
| | SNLI | 0.10 | 0.10 | 0.01 | 0.00 | 0.00 | 0.00 |
| | ANLI | 0.10 | 0.10 | 0.01 | 0.00 | 0.00 | 0.00 |

Table 4: Hyperparameters for fine-tuning on SNLI, MNLI, and ANLI.

| Model | Training Data | SNLI | MNLI | ANLI (R1) | ANLI (R2) | ANLI (R3) |
|---|---|---|---|---|---|---|
| BERT base | ANLI | 0.915 | 0.84 | 0.556 | 0.457 | 0.43 |
| | MNLI | 0.792 | 0.84 | 0.244 | 0.28 | 0.312 |
| | SNLI | 0.899 | 0.723 | 0.263 | 0.312 | 0.314 |
| BERT large | ANLI | 0.924 | 0.871 | 0.611 | 0.45 | 0.45 |
| | MNLI | 0.837 | 0.862 | 0.294 | 0.274 | 0.313 |
| | SNLI | 0.919 | 0.776 | 0.29 | 0.31 | 0.327 |
| DeBERTa base | ANLI | 0.936 | 0.909 | 0.73 | 0.505 | 0.511 |
| | MNLI | 0.892 | 0.902 | 0.475 | 0.329 | 0.342 |
| | SNLI | 0.933 | 0.842 | 0.39 | 0.328 | 0.335 |
| DeBERTa large | ANLI | 0.94 | 0.918 | 0.819 | 0.647 | 0.615 |
| | MNLI | 0.908 | 0.912 | 0.575 | 0.407 | 0.408 |
| | SNLI | 0.939 | 0.878 | 0.541 | 0.427 | 0.413 |
| RoBERTa base | ANLI | 0.844 | 0.879 | 0.59 | 0.446 | 0.406 |
| | MNLI | 0.841 | 0.878 | 0.302 | 0.308 | 0.27 |
| | SNLI | 0.91 | 0.788 | 0.321 | 0.325 | 0.316 |
| RoBERTa large | ANLI | 0.884 | 0.908 | 0.686 | 0.375 | 0.386 |
| | MNLI | 0.89 | 0.903 | 0.471 | 0.254 | 0.269 |
| | SNLI | 0.927 | 0.842 | 0.383 | 0.305 | 0.295 |

Table 5: The accuracy evaluated on standard dataset.

NLI fine-tuned models were initialized using pre-trained models and further fine-tuned on NLI datasets, implemented on the transformer package provided by Huggingface (Wolf et al. 2019). We fine-tune with Adam optimizer using the hyperparameters listed in table 4, where the training objective is the cross-entropy loss between the labels and the predictions. After fine-tuning, the models achieve promising accuracy in validation set, as shown in Table 5.

## Appendix B Details for Entailment Scores Collection

In this work, nearly 200 human annotators were involved in, resulting in 33975 pairwise entailment scores in total. For ChatGPT, we adopted the prompt with zero-shot setting as following:

CCL 2024

**Instruction**

Please judge whether the first sentence can deduce the second sentence in the following sentence pairs, please choose from the following four options:
A. cannot be inferred
B. low probability to be inferred
C. high probability to be inferred
D. can be inferred
First sentence: *A number of kids gets instruction from band instructor.*
Second sentence: *There is more than one person on the street.*

## Appendix C The Sentence Spaces of All Models

CCL 2024

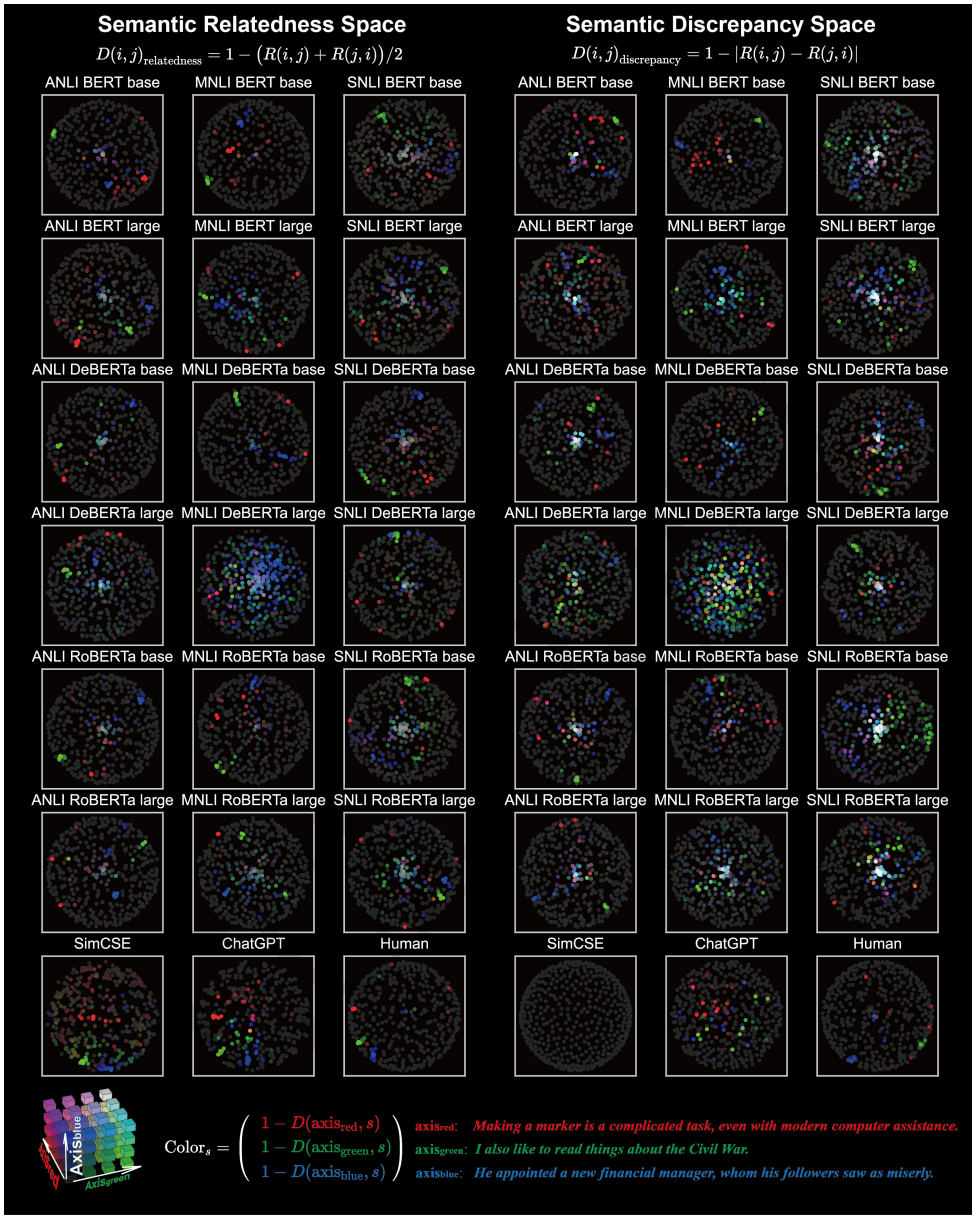



Figure 3: Visualization of the sentence space of semantic relatedness and semantic discrepancy respectively.

# AuditWen: An Open-Source Large Language Model for Audit

**Jiajia Huang**
Nanjing Audit University
Nanjing, Jiangsu, China
huangjj@nau.edu.cn

**Haoran Zhu**
Nanjing Audit University
Nanjing, Jiangsu, China
zhuhr@stu.nau.edu.cn

**Chao Xu**
Nanjing Audit University
Nanjing, Jiangsu, China
xuchao@nau.edu.cn

**Tianming Zhan**
Nanjing Audit University
Nanjing, Jiangsu, China
ztm@nau.edu.cn

**Qianqian Xie**
The FinAI
Singapore
xqq.sincere@gmail.com

**Jimin Huang**
The FinAI
Singapore
jimin@chancefocus.com

## Abstract

Intelligent auditing represents a crucial advancement in modern audit practices, enhancing both the quality and efficiency of audits within the realm of artificial intelligence. With the rise of large language model (LLM), there is enormous potential for intelligent models to contribute to audit domain. However, general LLMs applied in audit domain face the challenges of lacking specialized knowledge and the presence of data biases. To overcome these challenges, this study introduces AuditWen, an open-source audit LLM by fine-tuning Qwen with constructing instruction data from audit domain. We first outline the application scenarios for LLMs in the audit and extract requirements that shape the development of LLMs tailored for audit purposes. We then propose an audit LLM, called AuditWen, by fine-tuning Qwen with constructing 30k instruction dataset from 15 audit tasks and 3 layers. In evaluation stage, we proposed a benchmark with 5k instructions that covers a set of critical audit tasks derived from the application scenarios. With the benchmark, we compare AuditWen with other existing LLMs from information extraction, question answering and document generation. The experimental results demonstrate superior performance of AuditWen both in question understanding and answer generation, making it an immediately valuable tool for audit.

**Keyword** AuditWen, LLM, instruction dataset, fine-tuning, benchmark

## 1 Introduction

Audit is an independent economic supervision activity conducted by governmental agencies or a special organ in accordance with the law to conduct pre-and-post-event reviews of major projects and financial revenues and expenditures of financial institutions or enterprises. In recent years, with the development of big data, the data foundation and audit methodology of national audit are also undergoing changes (Zhang et al., 2020). The audit methodology is transitioning from big data audit to intelligent audit (Huang et al., 2023), aiming at recommending or selecting the optimal strategy for audit decision-making through the extensive integration of machine learning, deep learning, and other information technologies. With the emergence of ChatGPT [0], large language models (LLMs) (Che et al., 2023) have attracted much attention from researchers. Its smooth natural dialogue and document generation capabilities have rendered it widely used in various fields, such as in financial (Xie et al., 2023), medical (Singhal et al., 2023), legal (Dai et al., 2023) and so on. A large language model is a deep learning model with a very high number of parameters and computational power that can automatically learn the syntax, semantics, and context of input natural language and can generate text of corresponding to it. As a powerful artificial intelligence technology, large language model possess a strong capacity for understanding and generating natural language and can provide innovative solutions for the audit.

However, the current general LLMs commonly encounter issues like a deficiency in domain-specific knowledge and the existence of data bias. Similar to their application in other domain-oriented tasks,

©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

[0]https://chat.openai.com

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1351-1365, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

LLMs face challenges when directly applied to auditing, including difficulties in understanding input issues clearly and providing accurate responses to fact-based tasks, a phenomenon known as hallucination (Che et al., 2023). Moreover, auditors argue that intelligent auditing with LLMs should prioritize collaboration between individuals and the model to jointly accomplish complex audit tasks (Huang et al., 2023). This demand necessitates that LLMs not only comprehend concepts, entities, and knowledge within the audit domain, but also master the fundamental processes of audit work to assist auditors in achieving high-quality results. LLMs excel in context memory, knowledge retrieval, and text generation, thereby offering unique advantages in this regard.

Therefore, it is essential to train a LLM specifically for the audit domain, aligning with the actual requirements and raw data of auditing practices. By refining and tailoring LLM tasks to align with auditing requirements, the audit-focused LLM should grasp the terminology, concepts, and regulations of auditing, ultimately delivering more precise and dependable results, especially for the complicated audit tasks. Guided by the practical applications of national audit, this study aims to identify potential uses of LLM in the audit domain, collect high-quality audit-relevant raw texts and further construct an instruction dataset to build a large language model tailored for audit by fine-tuning a state-of-the-art LLM. This model is referred to as AuditWen.

The contributions of this study are as follows:

**(1) Scenarios abstraction**. We have categorized the application scenarios of LLM in audit as core requirements, regulatory requirements, and derived requirements. The abstracted scenarios can serve as a roadmap for future researchers to advance the development of LLMs for auditing purposes.

**(2) Multi-audit-tasks**. We abstract the corresponding NLP (natural language processing) tasks of LLM from 3 layers, including (a) phrase layer with information extraction and phrase classification, (b) sentence layer with audit-issue summary, audit legal recommendation and QA tasks, (c) document layer with audit risk analysis and audit report generation.

**(3) First open-source audit LLM**. It is the first open-source LLM for audit. We have openly released the AuditWen [1], including the instruction dataset, the evaluation benchmark and the model to encourage open research and transparency in the research field.

**(4) Outstanding performance**. AuditWen shows significant performance on various of audit NLP tasks compared with the state-of-the-art LLMs, especially in audit issue summary and legal recommendation. AuditWen can be directly used in some audit practice scenario.

## 2 Related Works

**Open Sourced Large Language Models**. The GPT (Generative Pre-Training) series of models released by OpenAI has ushered in a new era of large language model. GPTs and other LLMs demonstrate powerful language understanding and generation capabilities through pre-training on extensive text datasets followed by fine-tuning for diverse NLP tasks. Most of the open-source LLMs, such as LLaMA (Touvron et al., 2023), Alpaca (Taori et al., 2023) , Baichuan (Yang et al., 2023), ChatGLM[2], Qwen-VL Chat (Bai et al., 2023), have parameters ranging from 7B and 13B up to 65B. This rapid increase in the number of parameters results in notable enhancements in model power and performance, enabling LLMs to excel in NLP tasks. Generally, LLM building process consists of four main stages, i.e., pre-training, supervised fine-tuning (SFT), reward modeling and reinforcement learning from human feedback.Among the four stages, supervised fine-tuning of a base LLM with instruction dataset can produce superior answers to user queries compared to the base model, all at a lower cost. Along this line, some domain LLMs are proposed by constructing domain-oriented instruction dataset and fine-tuning base LLM (e.g,. LLaMA) with the dataset. For example, PIXIU (Xie et al., 2023) is an LLM specialized in financial domain, whereas HuaTuo (Wang et al., 2023) is tailored for the medical domain, both fine-tuned using LLaMA. However, there is currently a lack of open-source LLMs and instruction tuning data specifically tailored for auditing purposes.

[1]The AuditWen is available at : https://github.com/HooRin/AuditWen

[2]https://github.com/THUDM/ChatGLM-6B

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1351-1365, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

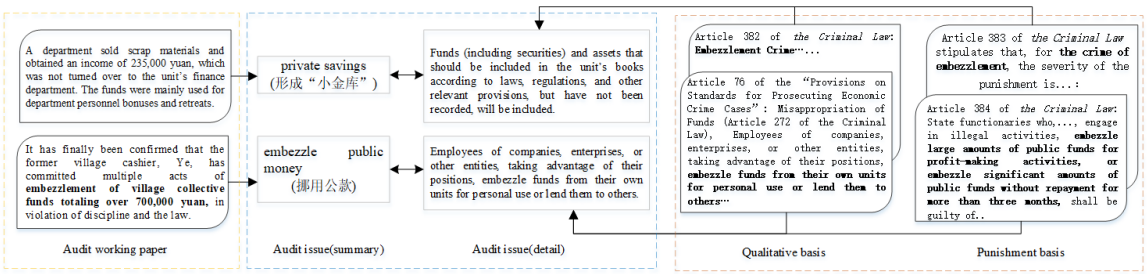



Figure 1: An audit issue schema derived from audit working paper and laws and regulations.

**LLM tasks and domain-oriented benchmarks**. To compare the performance of different LLMs, researchers have designed various types of LLM evaluation benchmarks and released evaluation reports (Cheng et al., 2023)(Guo et al., 2023). Among them, Microsoft Research Asia (Guo et al., 2023) has comprehensively sorted out and summarized 219 relevant studies from the perspectives of evaluation objects, evaluation fields and evaluation methods. In general, the current evaluation tasks are mainly designed from the perspectives of information extraction, text classification and text generation. The evaluation tasks of information extraction mainly include named entity recognition (NER) and key element recognition. The task of text classification includes emotion classification, text classification and entity classification. Text generation tasks include answer generation based on input question, machine translation, document generation in a specified form. Based on the above classification of evaluation tasks, researchers have released the open-sources of the domain evaluation benchmark datasets and fine-tuned domain large language models, such as PIXIU(Xie et al., 2023), FinBen (Xie et al., 2024), LAiW (Dai et al., 2023), HuaTuo (Wang et al., 2023) and so on.

Currently, there is no established benchmark for evaluating LLMs in the field of audit. According to the audit service requirements, this study designs 15 different LLM tasks across 3 layers, constructs the corresponding instruction datasets, and release multi-dimensional evaluation results for both existing mainstream LLMs and our fine-tuned audit-specific LLM, AuditWen.

## 3 Application Scenarios of LLM in Audit Domain

### 3.1 Audit issue summary and laws recommendation

The primary task of audit is to identify any potential audit issues within a project and determine which laws and regulations can serve as the audit basis. From this perspective, auditors are seeking LLMs to assist in summarizing audit issues based on audit working papers and recommending suitable laws and regulations as both qualitative and punishment basis.

The primary challenge in the application is that an internal auditor may have a divergent qualitative basis for an audit issue compared to a social auditor based on the case description in the audit working paper. For example, an internal auditor may use items from enterprise internal control manual as qualitative basis without any penalty provision, while a social auditor may refer to items in *Accounting Law and Criminal Law* for punishment. To address this challenge, we propose an audit issue schema that summarizes audit issue from case description and aligns them with the clauses of laws and regulations simultaneously, as shown in Figure 1. We hope to bridge a gap between the clause of laws and regulations and the audit issue.

### 3.2 Audit Relevant Question and Answer

The secondary task of LLM used in audit is to answer question related to audit, such as questions list in Table 1. These questions pertain to defining an audit concept, understanding the specifics of a particular clause of a law, determining the methods for investigating and verifying audit issues, and identifying the necessary data to be collected. These diverse questions prompt us to gather relevant audit documents pertaining to audit cases, audit criteria, audit guidelines, and so on. When assessing the quality of answers

| Query | Answer |
|---|---|
| What internal control information does the company need to disclose? (公司需要披露哪些内部控制信息？) | The Company shall fully disclose any internal control information that has a significant impact on investors' investment decisions.(凡对投资者投资决策有重大影响的内部控制信息，公司均应充分披露。) |
| What are the responsibilities of the audit institution under the Internal Audit Regulations? (内部审计条例规定的审计机关的职责有哪些？) | According to Article 23 of Chapter 5 of the Internal Audit Regulations of XX Province, the responsibilities of audit institutions include the following:... (根据XX省内部审计条例第五章第二十三条，审计机关的职责包含以下几项：...) |
| What methods should be adopted in the audit of significant secured loans to identify and conduct detailed investigation? (重大担保借款审计中需要采取哪些方法来识别并展开详细调查？) | Review meeting minutes and materials, documents, contract approval documents, etc., review decision-making process, guarantee scope, scale, etc. (调阅会议纪要及上会材料、文书、合同审批文件等，审核决策流程、担保范围、规模等。) |
| What key information should be collected when conducting a construction project performance audit? (在进行建设项目绩效审计时，应该收集哪些关键资料？) | When conducting a construction project performance audit, key data include: approved project proposals and feasibility study reports,... (在进行建设项目绩效审计时，关键资料包括：已获批准的项目建议书和可行性研究报告，...) |

Table 1: Examples of possible QA proposed by auditor.

| Id | Query |
|---|---|
| Q1 | Please extract entity about the audited organization from the following documents. (请从下面文档中抽取出被审计单位信息。) |
| Q2 | Please judge whether Company A is losing money according to the following statement.(请根据下面的报表判断A公司是否亏损？) |
| Q3 | Please write a business leader economic responsibility audit report template. (请撰写出一个企业领导人经济责任审计报告模板。) |
| Q4 | Based on the uploaded audit draft and the generated audit report template, please write the audit process and method of XX leader's accountability audit. (基于上传的审计底稿和生成的审计报告模板，请撰写XX领导人经责审计的审计过程与方法。) |
| Q5 | Based on the uploaded audit ledger list, audit ledgers belonging to the same audit issue are merged into the same document and output. (基于上传的审计台账列表，将属于同一审计问题的审计台账归并到同一文档中并输出。) |

Table 2: The potential tasks that may be assigned to an audit intelligent assistant.

generated by LLM, it is crucial to minimize the occurrence of hallucination responses and ensure the retrieval of original text based on existing system documents and other relevant content.

### 3.3 Audit assistant

Further derive requirement of LLM applied in audit domain is LLM can act as an intelligent assistant and help auditor to extract specified phrase from audit document, do accounting relevant numerical calculation, generate an outline for an audit report and further fill content based on the given audit working papers. The possible case questions are list in Table 2. Audit assistant usually need to execute fine-grained NLP task step by step, such as information extraction, multi-documents summarization and document generation.Additionally, audit assistants must achieve collaborative work between humans and machines with the guidance of human-provided knowledge.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1351-1365, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

| Entity tag | Description | Examples |
|---|---|---|
| audit issue（审计问题） | word or phrase of expressing an audit issue | 同一个人账户重复缴存, 规避招标，“小金库” |
| audit basis（审计依据） | word or phrase of expressing a law or regulation name | 招标投标法，中华人民共和国刑法，会计准则 |
| audit organization（审计对象） | entity of expressing an organization under audit | 国家机关、民办非企业单位，城市发展银行 |

Table 3: Audit entity types defined in audit domain.

## 4 AIT: Audit Instruction Dataset and Tuning

In this section, we initially outline the tasks of audit LLM based on the application scenarios of audit. Then we collect source data and design relevant instruction dataset and evaluation benchmark for audit LLM. At last, we build AuditWen by fine-tuning Qwen (Bai et al., 2023) with AIT.

### 4.1 Task abstraction for audit LLM

Based on the application scenarios of audit, we abstract the audit tasks from three levels, namely, sentence, paragraph and documents, as example shown in Table 4.

#### 4.1.1 Sentence level

This level focus on information extraction from sentence and phrase classification.

**Audit NER.** Accurately extract audit entity from text is the most elementary task for understanding audit content.We have developed an audit name entity recognition (NER) datasets from annotated sentences that include three types of entities, ORG, audit-issue and audit-basis, as shown in Table 3.

**Relation Classification.** Based on two audit entities extracted from a sentence, this task needs to predict the relation between the entity pair from given category set. The relations are defined in Table 7 in Appendix. This task can be used to expand audit knowledge graph by extracting information from unstructured text using LLM.

**Phrase classification.** Predict the category of an audit phrase from a set of options, where the phrase is (1) an audit-item entity that need to be classified into one of the given audit item type. (2) An audit issue relevant entity that need to be classified into one of the given audit type. (3) An law and regulation name that need to be classified into one of the given law and regulation category.

#### 4.1.2 Paragraph level

Question answer (QA) is the task of answering an audit question based on provided information, as shown in Table 1. In this level, we defined several types of question and answer tasks to make LLM understand the common question in audit.

**Definition of audit entity**, namely answer the definition of an audit entity, such as *what is internal audit?* The task makes LLM understand the concept and explanation of common audit entity.

**Audit-legal relevant question**, namely answer the question related to audit law, standards, guidelines. These part of QA pairs are very important for tuning an audit LLM, since the core scenario of audit LLM is to recommend appropriated laws and regulations as the audit basis for given audit issue.

**Audit-issue relevant question**, namely answer the question related to audit issue, including (1) use a phrase to summarize the audit issue based on case description, (2) describe the specific performance of an audit issue, (3) recommend appropriate laws for a given audit issue.

**Other-audit relevant question**. These QA pairs refer to (1) what method can be used in an audit case and what material need to prepare further, (2) what is the objective of an audit project, (3) list out the audit items of an audit project.

CCL 2024

#### 4.1.3 Documents level

This level focus on comprehensive documents analysis and generation, including audit risk/problem analysis, audit case/report generation, as shown in Q3-Q5 of Table 2.

**Risk/problem analysis**, namely analyzes the latent risks or issues of an audit project based on provided background information.

**Audit document generation**, namely generate an outline, or a template or a complete document based on input query, including (1) generate the audit process for a certain audit case, (2) outline the structure of an audit report for a specific audit matter.

### 4.2 Instruction dataset construction

Building upon the audit-oriented LLM tasks, we have developed an Audit Instruction Tuning dataset (AIT) specific to each task. Based on raw texts collected from audit domain discussed in Section 5.1, we need to construct a proper instruction for each of the raw texts.

First of all, for sentence level tasks and part of questions presented in paragraph level, we write five different instructions for each task and evaluate their performance on current LLM based on PIXIU project [3].Then the best instruction is saved for further constructing more instruction data. For audit-legal relevant question in paragraph level that concerns to items in audit laws, we used GPT-4 to generate a question and corresponding answer. For audit report generation task, we write one proper instruction for it because the query of this task is concise. AIT is the first large-scale instruction-tuning and evaluation benchmark dataset for audit LLMs that condensed from audit applications.

Generally, following the instructions proposed in PIXIU (Xie et al., 2023), we build instruction tuning samples with the following templates:

- Template (1) : [Task prompt] with {Context: [input text]}, [question] with {category}, Answer: [output]
- Template (2): [Task prompt] with Context: [input text] and [question], Answer:[output]

[task prompt] is the prompt designed for each type of the tasks, category used in classification tasks of sentence level to list out all categories, [input context] contains the input audit context of each task, such as a sentence or a paragraph. [question] is the final question or demand based on Context. [output] is the corresponding answer for the input text, such as the category in classification task or the truth answer in QA task. Examples of the instruction of each task is shown in Table 4.

### 4.3 Fine-tuning

We further build AuditWen by fine-tuning Qwen (Bai et al., 2023) with AIT because AIT is Chinese dataset and evaluation results on several LLMs show that Qwen achieves best performance on our evaluation benchmark dataset. To fine-tune the audit LLM, the audit instruction datasets outlined in Section 4.2 are divided into training, validation, and test sets. All the tasks in the training and validation sets are mixed together for fine-tuning, while each test set is utilized to evaluate the performance of AuditWen and other baseline LLMs.

We fine-tune Qwen-7B-chat[4] with 15 epochs based on AdamW optimizer (Loshchilov and Hutter, 2017). The batch size is set to 8, the initial learning rate is 3e-4, learning rate scheduler type choose as cosine, and warm up steps to 0.01. The AuditWen is fine-tuned on 8*A40 GPU with LoRA (Low-Rank Adaptation) (Hu et al., 2023) where the LoRA rank set to 64, LoRA alpha set to 16 and LoRA dropout set to 0.05. The maximum length of input texts is 2048. We choose LoRA for fine-tuning is because the method can make LLM achieve a good result in downstream task with training a few additional parameters. The addition parameter matrix merges with the large-scale of original parameters by reparametrization to form a new model for inference.

[3]PIXIU is available at: https://github.com/chancefocus/PIXIU

[4]The model of Qwen-7B-chat is downloaded from https://huggingface.co/Qwen/Qwen-7B-Chat/tree/main

| Task name | Template | Examples of an instruction data |
|---|---|---|
| Audit NER | T(1) | ”query”: 文本: 通过对证券公司和国有企业审计，...。”,”在上面的文本中，请完成命名实体识别任务，即识别代表审计疑点(’auditissue’)、机构(’ORG’)、审计法律法规(’auditlbasis’)三类实体类型的实体名称，答案应遵循格式:”实体名称, 实体类型”。<br>”answer”: ”证券公司, ORG”.<br>”label”: [”O”, ”O”, ”O”, ”B-ORG”, ”I-ORG”, ”I-ORG”, ”I-ORG”, ””,...] |
| Definition of audit entity | T(2) | ”query”: 请问什么是营业外收入?<br>”answer”: 该科目核算的是企业发生的与其生产经营无直接关系的各项收入，包括固定资产盘盈、处置固定资产净收益....。 |
| Audit-legal relevant quesiton | T(2) | ”query”: 如果某公司违反了中华人民共和国证券法第九十条关于征集股东权利的规定，将受到何种法律后果?<br>”answer”:根据中华人民共和国证券法第一百九十九条，该公司将被责令改正并给予警告...。 |
| Risk/problem analysis | T(2) | ”query”: 在国有企业经济责任审计，资产审计可能存在哪些审计风险?<br>”answer”: 资产审计可能存在如下风险点：（一）客户管理效率低，没有全面调研客户资质、信用状况并动态跟踪...。 |
| Audit document generation | | Referred in Table 8 |

Table 4: Examples of the instruction data used in LLM tuning dataset.

## 5 Experimental Results

### 5.1 Statistics of instruction dataset

To obtain domain data source for fine-tuning an audit LLM, we collect raw documents that relevant to definition of audit entity , audit relevant laws and kinds of structured audit cases that describe the detail process of an audit project, including audit issue, audit method, audit punish law and audit items. The raw data collected from baidubaike, public audit textbook, open law and other public website.

From the raw dataset, we construct an entity-relation classification dataset where two audit entities extracted from a given sentence and it's need to classify the relation between them from given category set. Here, the relations between an entity pair are defined in Table 7 in Appendix, and the truth category tag is labeled by human annotation. The rest of the classification tasks and entity extraction tasks are constructed with the similar way. Based on the raw classification task description and truth category tag, we converted each of them into instruction data with Template (1), as discussed in Section 4.2.

To construct audit-legal relevant instruction dataset, we gathered a substantial amount of audit-relevant laws, regulations, criterions and segmented each raw law or regulation into individual items. Then, GPT4 (OpenAI, 2023) is utilized to generate a question-answer pair (QA pair) based on the input items, as the examples shown in Table 8 in Appendix. The similarity between the original legal-item and the generated QA-pair are evaluated by BERT Score (F1) (Zhang et al., 2020). The similarity analysis reveals that over 80.1% of QA pairs exhibit a similarity score greater than 0.8, while 19% of QA pairs fall within the similarity range of 0.7 to 0.8,, which denotes that GTP4 can generate QA pair from given legal-item with high quality. Therefore, these QA-pairs can serve as instruction data that effectively capture the essence of the original legal content.

For the audit case/report generation task, we collected some representative audit cases or reports with various forms and convert each of them into an instruction data, where the query is a short instruction while the answer is a long document with given form. For the rest of the tasks in paragraph level, raw information are extracted from structured audit cases and converted into instruction data with Template (2) in accordance with specific conditions, as the examples shown in Table 4.

All of the train, validation and test sets for each of the tasks are shown in Table 5. For audit entity

CCL 2024

| Level | Task name | Sub-task name | #train/val./test | Annotation |
|---|---|---|---|---|
| Sentence level | | Audit NER | 4091/1022/1424 | human annotation |
| | | Relation classification | 817/232/117 | human annotation |
| | Phrase classification | audit entity cla. (AEC) | —/—/1578 | |
| | | audit-issue phrase cla. (AIC) | 1210/344/166 | human annotation |
| | | legal name cla. (LNC) | 1463/418/218 | |
| Paragraph level | Definition of audit entity | | 1756/500/19 | extract from raw text |
| | Audit-legal relevant question | | 15774/112/505 | generated by GPT-4 |
| | Audit issue | audit issue summary (AIS) | 253/71/36 | extract from raw text |
| | | audit issue describe (AID) | 202/56/29 | |
| | | legal recommendation (LR) | 1567/445/224 | |
| | Other-audit relevant question | audit procedures and material (APM) | 671/190/96 | |
| | | audit type and objectives (ATO) | 609/171/87 | |
| | | Other question (OQ) | 903/257/129 | |
| Documents level | | Audit case analysis | 544/151/77 | extract from raw text |
| | | Audit doc. generation | 48/11/6 | |
| | | Total | 29908/3980/4941 | |

Table 5: The details of our evaluation datasets. "Annotation" denotes the construction manner of the instruction data from raw data. source.

classification, only a test set is created to assess the generalization capability of AuditWen on untrained tasks. Therefore, 5-shot evaluation are employed for the task. In addition, as in the audit NER task, three new types of entities are defined that have not been encountered in base LLM, we also employ 5-shot prompting for evaluation. The rest of the tasks are evaluated under zero-shot prompting.

### 5.2 Evaluation of different LLMs

**Baseline Models**. Several strong and representative baseline models are selected to compare with our AuditWen model. For open-sources LLMs, Qwen-7B-Chat, ChatGLM3-6B are selected to perform zero-shot or 5-shot prompting on the audit evaluation benchmark dataset. For close-source LLM, GPT-4 (OpenAI, 2023) is selected.

**Evaluation Metrics**. As the tasks in sentence level are information extraction and classification, missing is employed to evaluate the proportion of prediction results that can be successfully inferred from LLM , while accuracy and F1 are employed to evaluate the classification effectiveness. As the tasks in paragraph level and document(s) level are Q&A task, BERT Score (F1) (Zhang et al., 2020), BART Score (Yuan et al., 2021) are employed to evaluated the similarity between the predict answer and the truth answer. For these two metrics, pre-train models with Chinese language are utilized, i.e., *bert-base-chinese* and *CPT* (Shao et al., 2021). In addition, we evaluate the definition of audit entity and legal recommendation with ROUGE (Lin and Hovy, 2003), because the answer of these tasks need to be more precise compared with other QA tasks. As word segmentation is a part of ROUGE evaluation, a user dictionary specific to the audit domain is created and loaded into the jieba segmentation tool. For the rest of the tasks, BERT Score (F1) and BART Score are used to evaluate the answer quality.

**Overall Performance**. From the 9 audit tasks evaluation results shown in Table 6, our fine-tuned model, AuditWen, significantly outperforms its base model QWen-7B-Chat and other state-of-the-art LLMs, especially in paragraph level and document level tasks. It is because fine-tuned the base LLM with domain-oriented instruction data enables the model to acquire domain-relevant knowledge, comprehend domain-specific queries, and generate outputs in the writing style typical of the audit domain.

CCL 2024

| Task name | Sub-task name | Metric | Qwen-7B-Chat | ChatGLM3-6B | GPT-4 | AuditWen |
|---|---|---|---|---|---|---|
| Audit NER | | Entity_F1 | 0.140 | 0.015 | 0.108 | **0.512** |
| Relation classification | | Accuracy | –/0.085* | 0.376/0.342* | 0.402/0.624* | **0.615**/0.188* |
| | | F1 | –/0.037* | 0.243/0.373* | 0.432/0.649* | **0.744**/0.220* |
| | | Missing | 0.410/0.00* | 0.008/0.000* | **0.000/0.000*** | 0.350/0.274* |
| Phrase classification | AEC | Accuracy | 0.716/0.763* | 0.493/0.540 | 0.679/**0.810*** | 0.601/0.720* |
| | | F1 | 0.710/0.734* | 0.583/0.612* | 0.697/**0.816*** | 0.612/0.716* |
| | | Missing | 0.042/0.00 | 0.146/0.000 | **0.023/0.000*** | 0.077/0.000* |
| | AIC | Accuracy | –/0.399* | 0.254/0.353* | 0.464/0.543* | 0.437/**0.601*** |
| | | F1 | –/0.347* | 0.193/0.252* | 0.484/0.557* | 0.428/**0.595*** |
| | | Missing | 0.751/0.000* | 0.078/0.058* | **0.000/0.000*** | 0.085/0.037* |
| | LNC | Accuracy | –/0.146* | 0.394/0.468* | 0.637/0.647* | **0.752**/0.431* |
| | | F1 | –/0.075* | 0.388/0.428* | 0.623/0.639* | **0.774**/0.405* |
| | | Missing | 0.766/0.165* | **0.000/0.000*** | 0.004/0.000* | 0.050/0.037* |
| Definition of audit entity | | ROUGE-1 | 0.245 | 0.22 | 0.202 | **0.298** |
| | | ROUGE-2 | 0.053 | 0.037 | 0.037 | **0.121** |
| | | ROUGE-L | 0.178 | 0.156 | 0.121 | **0.237** |
| | | BERT_Score | 0.678 | 0.670 | 0.662 | **0.702** |
| | | BART_Score | -4.527 | -4.535 | -4.391 | **-4.175** |
| Audit-legal relevant question | | BERT_Score | 0.696 | 0.671 | 0.665 | **0.723** |
| | | BART_Score | -3.659 | -3.356 | **-3.424** | -3.480 |
| Audit issue | AIS | BERT_Score | 0.634 | **0.644** | 0.634 | 0.642 |
| | | BART_Score | -4.470 | -4.485 | -4.524 | **-4.456** |
| | AID | BERT_Score | 0.696 | 0.674 | 0.655 | **0.792** |
| | | BART_Score | -4.048 | -3.827 | -3.996 | **-3.044** |
| | LR | ROUGE-1 | 0.247 | 0.268 | 0.275 | **0.530** |
| | | ROUGE-2 | 0.061 | 0.063 | 0.083 | **0.386** |
| | | ROUGE-L | 0.150 | 0.152 | 0.151 | **0.442** |
| | | BERT_Score | 0.654 | 0.665 | 0.677 | **0.785** |
| | | BART_Score | -4.799 | -4.192 | -3.661 | **-3.406** |
| Other-audit relevant question | APM | BERT_Score | 0.67 | 0.682 | 0.694 | **0.746** |
| | | BART_Score | -5.127 | -4.681 | -5.166 | **-4.514** |
| | ATO | BERT_Score | 0.723 | 0.697 | 0.634 | **0.907** |
| | | BART_Score | -3.794 | -3.650 | -4.069 | **-1.587** |
| | OQ | BERT_Score | 0.704 | 0.663 | 0.635 | **0.900** |
| | | BART_Score | -3.284 | -3.171 | -2.985 | **-1.202** |
| Audit case analysis | | BERT_Score | 0.67 | 0.678 | 0.667 | **0.84** |
| | | BART_Score | -4.854 | -3.61 | -3.291 | **-3.031** |
| Audit doc. generation | | BERT_Score | 0.658 | 0.668 | 0.670 | **0.684** |
| | | BART_Score | -5.584 | -5.003 | **-4.782** | -5.011 |

Table 6: The overall performance of different LLMs on audit evaluation benchmark, - denotes inadmissible inference result, * denotes the 5-shot evaluation result.

In the NER task, AuditWen demonstrates significantly higher entity F1 scores compared to baseline models in the 5-shot evaluation, indicating that baselines struggle to accurately identify named entities when provided with five examples from each category for inference.

In phrase classification tasks, including audit entity/ audit issue and legal name classification, AuditWen achieves competitive results compared to GPT-4, and outperforms the other models in F1 and accuracy,

CCL 2024

while ChatGLM3-6B and GPT-4 achieve much lower missing rate. Furthermore, comparing the the zero-shot evaluation results of QWen-7B-Chat and AuditWen across a range of phrase classification tasks, it is observed that QWen-7B-Chat may struggle in zero-shot inference due to a high missing rate, whereas AuditWen excels in overcoming this challenge and achieves higher accuracy.

Comparing the zero-shot and 5-shot result of different models, it is evident that baseline LLMs achieve higher accuracy and lower missing rates under the 5-shot setting, whereas AuditWen demonstrates higher accuracy under the zero-shot setting for relation classification and legal name classification(LNC). It denotes the model can be used for inference without providing extra samples, which further demonstrates the superior domain-generalization capabilities of AuditWen .

In the paragraph level and document tasks, AuditWen achieves much higher BERT Score and BART Score in legal recommendation, other-audit relevant question and risk/problem analysis. We believe that the success of AuditWen in these tasks is not only attributed to the suitable instruction template but also to the scale of the fine-tuning dataset for the task.

**Further Analysis**. We further analyze the influences of instruction template on the performance of LLMs on different tasks. For the tasks of relation classification and entity classification, varying instruction templates yield different results for Qwen-7B-Chat and ChatGLM3-6B. For instance, when the question is placed at the beginning of the template, Qwen-7B-chat exhibits poor performance on classification tasks, with accuracy rates of 0.019 for audit-issue phrase classification and 0.009 for legal name classification. ChatGLM3-6B also shows the similar result. When constructing instruction data using Template (1), ChatGLM3-6B achieves the highest accuracy and F1 score in audit entity classification, whereas it only attains an accuracy of 0.312 on the task when the question is placed at the beginning of the template. In addition, for relation classification and legal name classification(LNC), AuditWen shows better performance on zero-shot rather than on 5-shot. This may be attributed to the model's better utilization of general patterns and prior knowledge, as well as its reduced reliance on specific tasks or domains.

### 5.3 Case study

Audit-relevant document, such as audit report, is totally different from the document in general domain. Table 8.3 in Appendix compares the audit reported generated by AuditWen and GPT-4. The result demonstrates that the AuditWen can both understand the task with generating a report outline of satisfying the template and can generate more detail content of matching the outline, although only dozens of train instructions in audit case/report generation task for fine-tuning. On the contrary, base models without fine-tuning with the task fail to generate the audit report of meeting specified format.

## 6 Conclusion and Discussion

In this study, we presented AuditWen, the first audit-oriented open-source large language model. Along with the model, we also release the fine-tune model AuditWen and the evaluation benchmark dataset. Drawing from the discussion on application scenarios of LLM in audit, we have identified various audit tasks. Subsequently, we gather and construct a large-scale audit instruction dataset to fine-tune a domain-specific large language model tailored for audit tasks. The extensive evaluation results on the proposed benchmark dataset demonstrated the effectiveness of the AuditWen.

Nevertheless, while acknowledging the positive contribution of this study, we also recognize the following limitations. **Resource Constraints**. Due to time constraints, the scale of dataset for fine-tuning AuditWen is limited, which may not support for fine-tuning model with larger scales. **Model and Training Constrains**. We only presented the AuditWen models with 7B parameters. Due to computational and resource constraints, AuditWen models with 14B or 30B have not been released so far.

For the further work, more relevant source texts about audit cases and statute will be collected and more elaborate tasks such as audit-issue phrase extraction from clause of statute will be constructed. Based on these dataset and tasks, we devote to train a larger-scale of audit-oriented LLM.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1351-1365, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

## 7 Acknowledgements

The work was supported by the National Science Foundation of China (NSFC, No.61802194 and No.71972102) and Research Projects in Natural Sciences at Universities in Jiangsu Province (No.23KJB520015).

## References

Aiyuan Yang, Bin Xiao, Bingning Wang, et al. 2023. Baichuan 2: Open Large-scale Language Models, *arXiv preprint*, arXiv:2309.10305.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics, *In Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, 150–157.

Edward J. Hu, Yelong Shen, Phillip Wallis, et.al. 2023. LoRA: Low-Rank Adaptation of Large Language Models, *arXiv preprint*, arXiv:2106.09685.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, et al. 2023. LLaMA: Open and Efficient Foundation Language Models, *arXiv preprint*, arXiv:2302.13971.

Haochun Wang, Chi Liu, et al. 2023. HuaTuo: Tuning LLaMA Model with Chinese Medical Knowledg, *arXiv preprint*, arXiv:2304.06975.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled Weight Decay Regularization, *arXiv preprint*, arXiv:1711.05101.

Jiajia Huang, Pengwei Li, Chao Xu. 2023. Research on the Mindset Reform and the Audit Platform Construction of Intelligent Audit *Auditing Research*, 2023,(05):11-20.

Jinze Bai, Shuai Bai, Shusheng Yang, et al. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond *arXiv preprint*,arXiv:2308.12966.

Karan Singhal, Shekoofeh Azizi, Tao Tu, et.al. 2023. Large Language Models Encode Clinical Knowledge *Nature* 620, 172–180 (2023)

Yongfu Dai, Duanyu Feng, Jimin Huang, et.al. 2023. LAiW: A Chinese Legal Large Language Models Benchmark. *arXiv preprint* arXiv: 2310.05620

OpenAI. 2023. GPT-4 Technical Report *arXiv preprint* arXiv: 2303.08774

Qianqian Xie, Weiguang Han, Xiao Zhang, et.al. 2023. PIXIU: A Large Language Model, Instruction Data and Evaluation Benchmark for Finance *The 37th Conference on Neural Information Processing Systems (NeurIPS 2023)*

Qianqian Xie, Weiguang Han, Zhengyu Chen, et.al. 2024. The FinBen: An Holistic Financial Benchmark for Large Language Models *arXiv preprint* arXiv:2402.12659

Qianglong Zhang, Chunyu Xing, et.al. 2020. Next-generation Internal Audit: Digitized and Intelligent *Auditing Research*, 2020(05):113-121.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, et al. 2023. Alpaca: A Strong, Replicable Instruction-Following Model *arXiv preprint*.

Tianyi Zhang, Varsha Kishore, Felix Wu, et.al. 2020. BERTScore: Evaluating Text Generation with BERT *In International Conference on Learning Representations*.

Wanxiang Che, Zhicheng Dou, Yansong Feng, et.al. 2023. Towards a Comprehensive Understanding of the Impact of Large Language Models on Natural Language Processing: Challenges, Opportunities and Future Directions. *Science China*, 2023, 53(09):1645-1687.

Weizhe Yuan, Graham Neubig, Pengfei Liu. 2021. BARTSCORE: Evaluating Generated Text as Text Generation. *The 35th Conference on Neural Information Processing Systems (NeurIPS 2021)*, 27263-27277

Yunfan Shao, Zhichao Geng, Yitao Liu et.al. 2021. CPT: A Pre-Trained Unbalanced Transformer for Both Chinese Language Understanding and Generation. *arXiv preprint*, arXiv:2109.05729.

CCL 2024

Yupeng Cheng, Xu Wang, Jindong Wang, et.al. 2023. A Survey on Evaluation of Large Language Models. *ACM Transactions on Intelligent Systems and Technology*, 2023.

Zishan Guo, Renren Jin, Chuang Liu et.al. 2023. Evaluating Large Language Models: A Comprehensive Survey. *arXiv preprint*, arXiv:2310.19736.

# 8 C EXAMPLES

## 8.1 C1. Relations defined between entity pairs and corresponding examples

We provide the relation define between two audit entities and shows an example of entity pair extraction from a sentence and their define relation in Table 7.

## 8.2 Examples of audit-legal relevant question generated by GPT-4.

We provide some examples of question-answer pair (QA pair) generated by GPT4 based on the input law item in Table 8.

## 8.3 Examples of audit-report generated by different LLMs.

We present a audit report generated by AuditWen and GPT-4 respectively. Table 9.

CCL 2024

| Relation name | Description | Entity pair | Text |
|---|---|---|---|
| fraud_of_audit | Relation between an audit item and its audit fraud | [住房公积金归集, 同一个人账户重复缴存] | 本词条介绍了住房公积金缴纳对象在住房公积金归集方面存在主要弊错，主要包括住房公积金缴纳对象同一个人账户重复缴存的情况. |
| item_of_audit | Relation between an audit instance and specific audit items | [证券公司负债业务，资产负债表] | 证券公司负债业务发生后，都要通过相应的会计科目反应和核算，最终表现为资产负债表上的的负债项目，达到负债的动态业务和静态业务反应相统一。 |
| law_of_audit | Relation between an audit issue and the corresponding law entity | [规避招标,招标投标法] | 《招标投标法》规定：招标方不得以任何方式将应招标的项目而不招标或将必须进行招标的项目化整为零或者以其他任何方式规避招标。 |
| method_of_audit | Relation between an audit item and the corresponding audit method entity | [合同履行情况审计，检查] | 合同履行情况审计是指对公共工程实施过程中的造价、质量、进度、安全、环境保护和水土保持等合同约定内容的执行结果进行检查。 |
| org_of_audit | Relation between an audit item and the corresponding audit unit | [国家机关, 住房公积金] [国有企业, 住房公积金]... | 国家机关、国有企业、城镇集体企业、外商投资企业、城镇私营企业及其他城镇企业、事业单位、民办非企业单位、社会团体（以下统称单位）及其在职职工，应当按《住房公积金管理条例》（国务院令第350号，以下简称《条例》）的规定缴存住房公积金。 |
| achievement_of_audit | Relation between an audit item and the corresponding audit achievement entity | [政府预算审计，审计报告] | 2003年审计署首次公开政府预算审计报告，政府预算审计逐步进入社会公众及媒体的视野。 |
| audited_of_org | Relation between an audit item and the corresponding audited unit | [保障性安居工程跟踪审计, 哈尔滨特派办] | 在2012年城镇保障性安居工程跟踪审计过程中，哈尔滨特派办应用联网审计数据平台进行审计并取得了较好的审计成果。 |
| included_domain | Relation between an audit item and its belonging domain | [污染减排审计，电力行业] | 本文介绍了持续推进电力行业污染减排审计过程和关键特点。 |

Table 7: Relations defined between entities.

CCL 2024

| Legal Name | Content of Item | QA generated by GPT-4 |
|---|---|---|
| 《第2205号内部审计具体准则——经济责任审计》(第一章第五条) | 本准则适用于党政工作部门、纪检监察机关、法院、检察院、事业单位和人民团体，国有及国有资本占控股地位或主导地位的企业（含金融机构）等单位的内部审计机构、内部审计人员所从事的经济责任审计活动，其他类型单位可以参照执行。 | Question：《第2205号内部审计具体准则——经济责任审计》适用于哪些类型的单位进行经济责任审计活动?<br>Answer：该准则适用于党政工作部门、纪检监察机关、法院、检察院、事业单位和人民团体，以及国有及国有资本占控股地位或主导地位的企业（包括金融机构）等单位的内部审计机构和内部审计人员所进行的经济责任审计活动。其他类型的单位也可以参照该准则执行经济责任审计。 |
| 中华人民共和国证券法(第十三章第一百九十九条) | 违反本法第九十条的规定征集股东权利的，责令改正，给予警告，可以处五十万元以下的罚款。 | Question：如果某公司违反了中华人民共和国证券法第九十条关于征集股东权利的规定，将受到何种法律后果?<br>Answer：根据中华人民共和国证券法第一百九十九条，该公司将被责令改正并给予警告，同时可能会被处以不超过五十万元的罚款。 |
| 中华人民共和国银行业监督管理法(第四章第三十九条) | 银行业金融机构有违法经营、经营管理不善等情形，不予撤销将严重危害金融秩序、损害公众利益的，国务院银行业监督管理机构有权予以撤销。 | Question：根据中华人民共和国银行业监督管理法，国务院银行业监督管理机构在什么情况下有权撤销银行业金融机构?<br>Answer：根据中华人民共和国银行业监督管理法第四章第三十九条的规定，当银行业金融机构存在违法经营或经营管理不善等情形，且这些情形不予撤销将严重危害金融秩序、损害公众利益时，国务院银行业监督管理机构有权对该金融机构予以撤销。 |
| 中华人民共和国招标投标法(第二章第十条) | 招标分为公开招标和邀请招标。公开招标，是指招标人以招标公告的方式邀请不特定的法人或者其他组织投标。邀请招标，是指招标人以投标邀请书的方式邀请特定的法人或者其他组织投标。 | Question：中华人民共和国招标投标法规定了哪两种招标方式?<br>Answer：中华人民共和国招标投标法规定了公开招标和邀请招标两种方式。公开招标是指招标人通过发布招标公告的方式邀请不特定法人或其他组织投标；邀请招标是指招标人通过发送投标邀请书的方式邀请特定的法人或其他组织投标。 |

Table 8: Relations defined between entities.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1351-1365, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

| | |
|---|---|
| **Prompt** | Please write a template for the leadership debriefing reprot in internal audit. (请撰写一份内部审计中领导人述职报告模板. ) |
| **Ground truth** | ***〔姓名〕同志述职报告<br>一、基本情况〔说明：本部分主要反映被审计领导干部任职期限、职责范围和分管的工作。〕<br>二、主要业务工作开展情况〔说明：本部分主要包括：1.任期内贯彻执行党和国家有关经济方针政策和决策部署、推动所在单位可持续发展情况；2.任期内重大经济决策事项、决策过程及其执行效果；3.任期内各项工作任务目标完成情况；4.重要规章制度及内部控制制度的制定、完善和执行情况；5.任职前和任期内重大经济遗留问题及其处理情况等内容。〕<br>三、党风廉政建设情况〔说明：包括任期内履行推动所在单位党风廉政建设职责和个人遵守廉洁从业规定的情况。〕<br>四、存在的不足和需要改进的方面。<br>五、其他需要说明的情况<br>述职人：***〔述职人本人签字〕****年**月**日 |
| **AuditWen** | 领导人员述职报告<br>一、基本情况介绍（主要包括：任职时间、分工情况、分管工作等）<br>二、履行职责情况<br>（一）坚持政治统领，增强贯彻执行党的路线方针政策和决策部署的坚定性：1．（简要叙述在贯彻落实上级部门重大政策措施、公司党组（企业党委）重要决策部署方面的情况）2．（简要叙述如何加强党建考核，推动各级单位落实党建工作责任的情况）3．（简要叙述推进党内监督，落实党风廉政建设责任的情况）<br>（二）强化管理，切实促进公司高质量发展4．（简要叙述是否及时制定或修改生产经营目标，并加强目标管理的情况）5．（简要叙述是否建立或完善生产经营管理制度，确保制度之间相互衔接、配套协作的情况）6．（简要叙述领导并督促解决生产经营中存在的突出问题，保障生产经营工作顺利进行的情况）<br>（三）狠抓 建设，注重队伍建设管理7．（简要叙述在组织培养选拔干部方面所做的事情）8．（简要叙述人才规划制定和人才培养方面所做的事情）9．（简要叙述加强团队建设，打造优秀集体的情况）10．（简要叙述推进企业文化建设，传播积极向上的企业文化的情况）<br>（四）严格遵守规章制度，树立良好作风11．（简要叙述是否严格遵守财经法规和规章制度的情况）12．（简要叙述履行“一岗双责”情况，是否既抓好经济工作又抓好廉政工作的情况）13．（简要叙述接受职工群众举报和批评意见的情况）<br>三、自我评价（简要回顾在任期内的工作表现，包括成绩和需要改进之处）<br>四、具体建议和要求（对于后续的工作提出建议和要求，例如：加强哪方面的管理工作，如何提升干部员工的积极性等） |
| **GPT-4** | 内部审计组长述职报告模板尊敬的[公司名称]领导及审计委员会：我谨代表内部审计组向大家报告[报告年份]年度的工作情况，并对即将开展的工作计划进行说明。<br>一、基本情况1. 组织结构- 组长：- 成员：- 分配职责：2. 审计范围- 覆盖部门：- 业务流程：- 财务报表：<br>二、年度工作回顾1. 主要成就- 完成审计项目数量：- 关键发现：- 实施改进措施：2. 风险管理- 识别关键风险：- 风险应对措施：3. 合规性审查- 检查遵守情况：- 发现违规行为：- 整改情况：4. 效率与效果- 优化流程与节约成本：- 提升内控质量：5. 职业发展- 团队培训与个人提升：- 资质认证：<br>三、存在问题及原因1. 问题概述2. 影响分析3. 原因分析<br>四、改进措施与计划1. 改进目标2. 实施计划- 短期措施：- 中长期措施：3. 监督与评估- 监督机制：- 评估指标：<br>五、下一阶段工作展望1. 审计计划- 重点审计项目：- 预计时间表：2. 资源配置- 人力资源：- 技术投入：3. 创新发展- 新技术应用：- 流程改革：<br>此致敬礼！[组长签名] [日期] |

Table 9: Audit report template generated by different LLMs.

# Chinese Grammatical Error Correction via Large Language Model Guided Optimization Training

**Xiao Liu**[1,2], **Ying Li**[1,2]✉, **Zhengtao Yu**[1,2]
[1]Faculty of Information Engineering and Automation,
Kunming University of Science and Technology, Kunming, China
[2]Yunnan Key Laboratory of Artificial Intelligence, Kunming, China
{sakuracedia,yingli_hlt}@foxmail.com, ztyu@hotmail.com

## Abstract

Pre-trained language model-based methods for Chinese Grammatical Error Correction (CGEC) are categorized into Seq2Seq and Seq2Edit types. However, both Seq2Seq and Seq2Edit models depend on high-quality training data significantly. Considering the strong generation and inference ability of large language models (LLMs), we propose a large language model-guided optimization training method to exploit LLMs to extract error knowledge to optimize the traditional CGEC model training process. On the one hand, we use error types and confusion sets as extra knowledge to guide LLMs to generate diverse pseudo data, thus extending the error distribution of our training data. On the other hand, LLMs are utilized to infer the predicted results from our CGEC models and obtain the re-training data, thus iteratively optimizing our pre-trained CGEC models. Experiments on two benchmark datasets show that our LLMs-guided optimization method with small-scale training data can achieve comparable results with baseline models with large-scale training data. Detailed comparison experiments demonstrate that both the early deviser pseudo data and the later re-training data are extremely useful for traditional CGEC model optimization training, and can benefit from each other. We will release our code and prompts at https://github.com/SakuraAcedia/llm-cgec-got to facilitate future work.

## 1 Introduction

Grammatical Error Correction (GEC) aims to detect and correct any possible grammatical errors in a given sentence (Wang et al., 2021), i.e., missing, redundant, substitution, and word-order four error types (Zhang et al., 2022a). As shown in Figure 1, the traditional GEC model receives a sentence containing errors and generates its corrected one. The GEC results have widely applied to various artificial intelligence tasks such as writing scenarios, speech recognition fields, and information retrieval scenes (Bryant et al., 2022; Zhang et al., 2023; Wang et al., 2023a).

Current typical GEC methods can be categorized into two lines, i.e., Seq2Seq models and Seq2Edit models. Seq2Seq models treat GEC as a generation task, which takes an erroneous sentence as input and obtains its corrected one autoregressively (Zhang et al., 2022b; Li et al., 2023a). Seq2Edit models regard GEC as a token classification task, which first predicts a sequence of text-editing operations and then generates target tokens based on the sequence operation labels (Zhang et al., 2022a; Omelianchuk et al., 2020). With the advance in pre-trained language models, both Seq2Seq and Seq2Edit models have achieved obvious improvements in Chinese grammatical error correction. However, their performances are highly dependent on the quality and scale of the training corpus (Bryant et al., 2022).

In the past few years, different data augmentation techniques have been extensively studied (Koyama et al., 2021; Stahlberg and Kumar, 2021; Ye et al., 2023). On the one hand, researchers attempt to decline the error distributions between synthetic and reality corpus by back-translation or error pattern constraint (Choe et al., 2019; Stahlberg and Kumar, 2021). On the other hand, some works exploit noise injection or MixEdit approaches to increase the diversity of synthetic corpus (Koyama et al., 2021; Ye et

©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License
✉ Corresponding author.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1366-1380, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

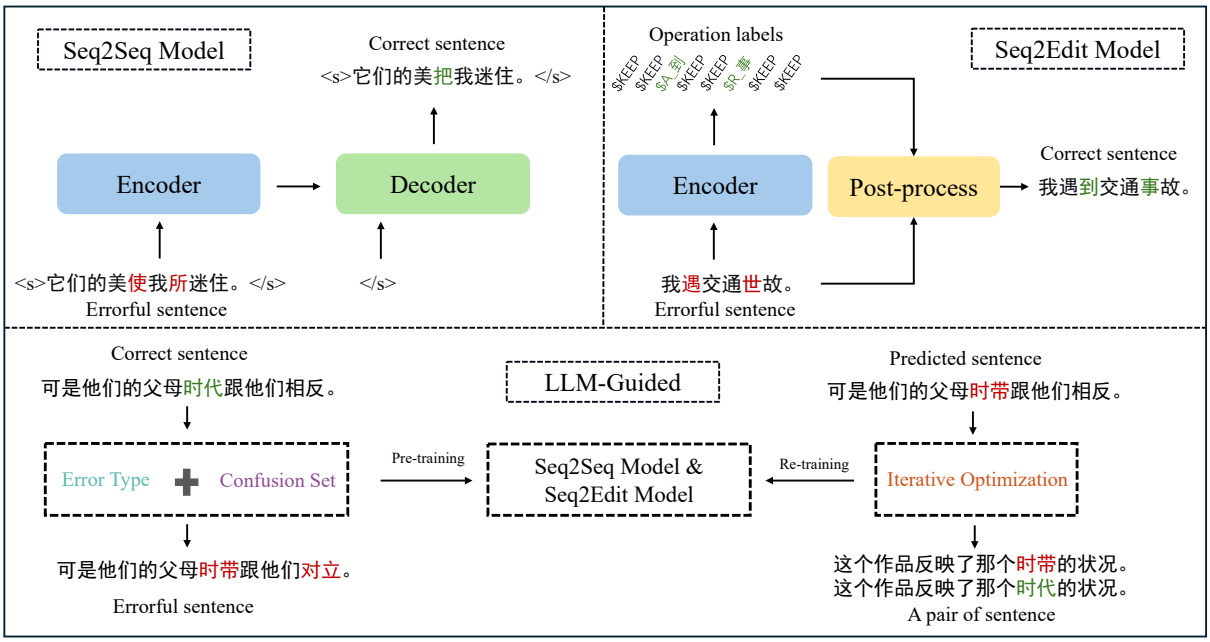



Figure 1: An example for CGEC corpus generation with LLMs.

al., 2023). More recently, large language models (LLMs) have demonstrated breakthrough performance in various natural language processing tasks (Wu et al., 2023; Fang et al., 2023). The LLMs benefit from a vast training corpus, providing them with ample grammatical knowledge, exceptional semantic comprehension, diversity generation, and strong representation capabilities (Brown et al., 2020; Zhou et al., 2023a). Several studies have investigated the possibility of LLMs for CGEC (Li et al., 2023b; Kaneko and Okazaki, 2023b). Research findings indicate that despite receiving supervised fine-tuning using CGEC data, the performance of LLMs on the CGEC task continues to be unsatisfactory (Zhang et al., 2023; Wu et al., 2023). The result may be due to the serious issue of overcorrection in LLMs. Recent research trends indicate that more comprehensive grammar information can be derived from LLMs (Song et al., 2023; Kaneko and Okazaki, 2023a). This grammar information can guide the GEC model to correct grammatical errors, enabling the model to correct complex errors that it was previously unable to correct. Due to the serious hallucination problem in LLMs, the corpus generated by LLMs may have slight deviations from real corpus. Therefore, simulating the distribution of grammatical errors in real-world scenarios and constructing high-quality training data remains a significant challenge.

To address this issue, we propose a large language model-guided optimization training method to enhance traditional GEC models with knowledge acquired from the LLMs. First, we utilize error types and the confusion set as extra knowledge to construct prompt instructions based on our given correct sentences, thus guiding LLMs to generate a diverse original synthetic corpus. Second, we directly fuse a few synthetic and reality corpora for the traditional GEC model training. Third, we utilize LLMs to analyze predicted sentences from GEC models, inferring whether there are any Chinese grammatical errors in the sentence that have not been corrected. Finally, we exploit LLMs to generate similar sentences based on the analysis results, which are used for re-training our GEC models iteratively, thus further enhancing the performance of GEC models.

Experiments on two benchmark datasets show that our proposed method achieves significant enhancements compared with the strong baselines, demonstrating that LLMs is extremely useful for optimizing the traditional GEC model training. Our thorough investigation highlights that error types and confusion sets are important factors in guiding LLMs to generate a diverse synthetic corpus, thus further expanding the error distribution of our training data. Detail comparison experiments show that re-training data generated by LLMs based on the inference and analysis results are essential for optimizing GEC model training since its error distributions are very similar to our testing data. The ablation study indicates

that both the early pseudo-corpus generation and the later iterative optimization are equally important in enhancing the performance of traditional GEC models, and they can benefit from each other.

## 2 Related Work

Grammatical error correction has been a challenging task in both natural language processing and machine learning fields. Recently, grammatical error correction has achieved significant improvement in English and Chinese. Now, we try to give a brief review of some representation approaches from two aspects, i.e., pre-trained language model-based methods and large language model-based methods.

**Pre-trained language model-based methods.** Motivated by the successful application of pre-trained language models on various artificial intelligence tasks, most previous works focus on enhancing the grammatical error correction performance by fusing the representations from pre-trained language models, which are mainly categorized into Seq2Seq and Seq2Edit two paradigms.

Seq2Seq models utilize encoder-decoder architectures to address the GEC task, which is drawn inspiration from neural machine translation (Kaneko et al., 2020; Zhang et al., 2022b; Zhou et al., 2023b). Later, Kaneko et al. (2020) extend the integration of pre-trained knowledge into the encoder-decoder models. Zhang et al. (2022b) utilizes syntactic information to enhance the model's performance. Zhou et al. (2023b) employ an external critic to appraise the suitability of the token to be produced gradually, impacting the selection of the succeeding token dynamically. For Seq2Edit models, they treat GEC as a sequence tagging task and predict token-level edit operations to be implemented on the input sentence (Omelianchuk et al., 2020; Zhang et al., 2022a). GECToR first iteratively annotates and revises the modified sequence, and then proceeds to correct it once more, thereby improving the speed of inference (Omelianchuk et al., 2020). Zhang et al. (2022a) integrate the GECToR model into the CGEC task and explore the impact of different pre-trained language models on this paradigm.

Compared with the Seq2Seq model, the Seq2Edit model has more efficient inference and robust error detection capabilities, but its performance has slightly declined (Li et al., 2022). In this work, we will give a detailed analysis of the influence of different training corpus on both Seq2Seq and Seq2Edit models.

**Large language model-based methods.** Recent researches on large language models indicate that increasing the size of language models can result in notable enhancements in performance across various downstream tasks (Kaplan et al., 2020). These LLMs have shown exceptional skill in adhering to and carrying out a variety of user instructions, leading to extensive acceptance among users who frequently engage with these models and utilize them for professional purposes (Brown et al., 2020; Dong et al., 2023; Li et al., 2023c; Zhao et al., 2023). Recently, there has been significant progress in corpus construction focused on reducing the manual annotation workload and automatically enhancing the scale, diversity, and creativity of corpus (Wang et al., 2023b; Asai et al., 2023; Zhou et al., 2023a). Wang et al. (2023b) provide an almost annotation-free method for aligning pre-trained language models with instructions. Asai et al. (2023) encourage LLMs to participate in self-reflective retrieval enhances the generation process, leading to improved quality and accuracy of language model output. Zhou et al. (2023a) design a small dataset containing 1000 samples that is used for training with standard supervised loss and demonstrate the potential of fine-tuning the LLMs on a high-quality small-scale dataset.

Recent work has explored the performance of various LLMs on the CGEC task (Fang et al., 2023; Qu and Wu, 2023; Fan et al., 2023; Zhang et al., 2023; Kaneko and Okazaki, 2023b; Song et al., 2023). Fang et al. (2023) evaluate the performance of ChatGPT on the CGEC task through in-context learning, highlighting its ability to produce coherent sentences while also exposing its tendency for excessive correction. Fan et al. (2023) employs ChatGPT to create ungrammatical sentences by incorporating specific clues. In cases where ungrammatical sentences lack clues, the author manually gathers such sentences from publicly accessible websites and rectifies them. Zhang et al. (2023) investigate the potential benefits of LLMs and enhance a writing-assistant scenario through multi-task instruction tuning. Despite supervised fine-tuning, the performance of LLMs remains unsatisfactory, indicating that they have not yet fully adapted to the GEC field. Consequently, researchers have started exploring if LLMs can play different roles in the GEC task, apart from merely acting as the corrector. Kaneko and Okazaki (2023b) propose a method to enhance the performance of the GEC task by leveraging LLMs for predicting edit

CCL 2024

spans. Song et al. (2023) introduce a novel concept of elucidating grammar mistakes and demonstrate the capability of LLMs to provide explanations for grammatical errors.

In a word, although the notable advancements achieved by typical Seq2Seq and Seq2Edit methods in the task of CGEC, their effectiveness is limited by the quality and scale of the training corpus. Considering the strong capability of generation and inference demonstrated by LLMs, we will attempt to utilize LLMs to simulate the distribution of realistic errors and further optimize the typical Seq2Seq and Seq2Edit model training.

## 3 Our Proposed Approach

Large language model has shown its strong capability of generation and inference in various natural language processing tasks, such as question answer, summary generation, and knowledge inference (Brown et al., 2020; Zhou et al., 2023a; Zhang et al., 2023). However, since the parameters for LLMs are huge, it is difficult to fine-tune LLMs to adapt to each specific natural language processing task. Hence, the utilization of the ability of LLMs to enhance the performance of traditional GEC models is still a key challenge for Chinese grammatical error correction. To address this issue, we propose a large language model-guided optimization training method. On the one hand, we utilize the LLMs to generate a small-scale diverse pseudo corpus automatically based on error types and confusion sets, thus enlarging the error distributions of training data. On the other hand, we exploit the inference ability of LLMs to analyze the predicted results from the traditional GEC models, thus obtaining a more high-quality re-training corpus to make up for the limitations in traditional model training.

As depicted in Figure 2, our proposed method mainly contains three stages, i.e., *pseudo corpus generation*, *traditional model training*, and *iterative optimization*. First, we combine the error types, correct sentences, and an additional confusion set in one prompt to guide the LLMs in creating sentences with different grammatical errors. Second, the generated pseudo corpus is used to enlarge the original training data for traditional GEC model training. Third, we iteratively utilize LLMs to analyze predicted sentences by GEC models, inferring whether there are any Chinese grammar errors in the sentence that have not been corrected. Finally, we use LLMs to generate similar sentences based on the analysis results to retrain GEC models, further enhancing the model performances.

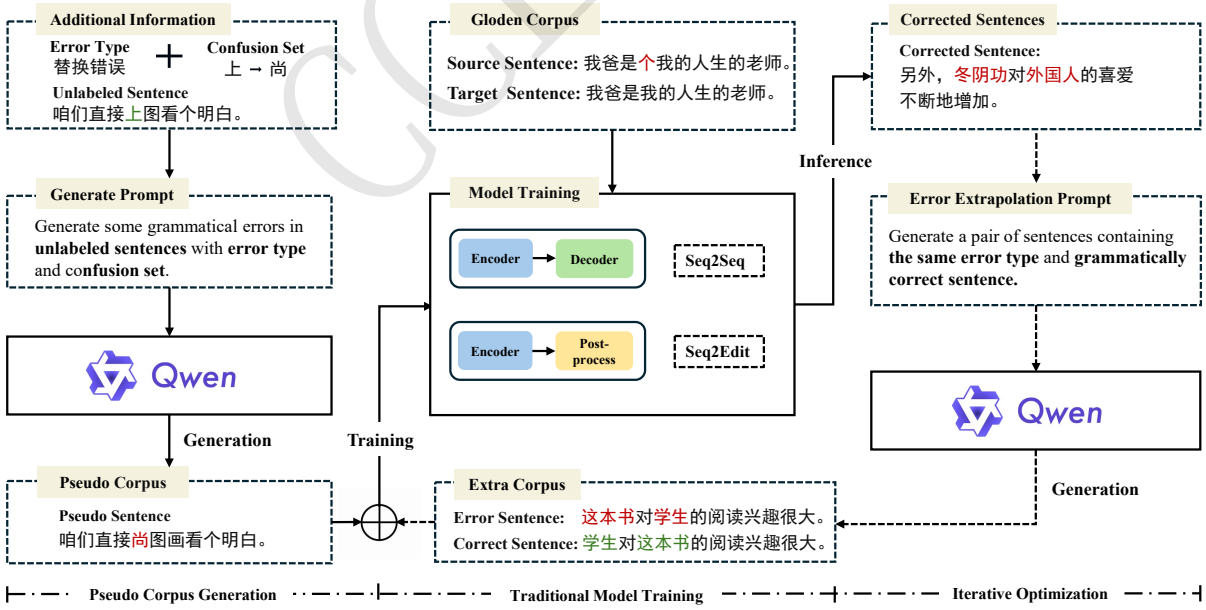



Figure 2: Framework of our proposed method.

### 3.1 Pseudo Corpus Generation

Traditional grammatical error correction models always rely on a large amount of high-quality parallel corpus to ensure effective performances. Therefore, how to automatically generate a high-quality corpus becomes an important factor in determining the model performance. Considering the strong generation capability of LLMs, we incorporate four error types and confusion set information as extra knowledge to constrain LLMs to generate a diverse pseudo corpus that covers all error types. In the practical experiments, we opt for the top 5 candidate characters that match each position within unlabeled sentences to maximize the utilization of the confusion set information. The process involves iterating through these candidates based on their probability in descending order. This selected candidate is then deemed as the final answer. The input of LLMs is formulated by converting error type $E$, confusion set information $C$, and unlabeled sentence $Y$ into an instruction template $\mathcal{T}_{\text{gen}}(E, C, Y)$:

$$\mathcal{T}_{\text{gen}}(E, C, Y) = \{\overbrace{z_1, ..., z_{i-1}}^{instruction}, \overbrace{z_i, ..., z_{j-1}}^{E}, \overbrace{z_j, ..., z_k}^{C} \overbrace{z_{k+1}, ..., z_{k+n}}^{Y}\}. \tag{1}$$

Specifically, we first encourage LLMs to analyze grammatical details such as punctuation, vocabulary, and sentence structure in the correct sentences. Then, we utilize LLMs to accurately analyze the part-of-speech of each word in a given sentence. Next, specific error operations are performed in this sentence based on the analyzed part-of-speech and confusion set information. For instance, when encountering a substitution error, we will substitute partial words in the sentence according to the part-of-speech and confusion set information. Finally, LLMs are employed to identify Chinese grammar errors related to punctuation, vocabulary, and sentence structure, thereby verifying the presence of grammatical errors in the generated sentences. Through the above process, we restrict the generation of pseudo-corpus by LLMs using error types and confusion set information. This "constrain-then-generate" approach enables us to generate more diverse, informative, and realistic pseudo instances.

### 3.2 Traditional Model Training

Given a sentence $X = \{x_1, x_2, \ldots, x_m\}$ that may include grammatical errors, the CGEC task aims to identify and correct the potential grammatical errors in $X$ and outputs the corresponding error-free sentence $Y = \{y_1, y_2, \ldots, y_n\}$.

**Seq2Seq Model.** The Seq2Seq model employs the encoder-decoder framework to generate the correct sentence based on the ungrammatical input sentence. It uses a multi-layer self-attention as the encoder module to encode the entire erroneous sentence $x_1, x_2, \ldots, x_m$ into the corresponding hidden states $\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_m$. Then, another multi-layer cross-attention is used as the decoder module to generate each token in $y_1, y_2, \ldots, y_n$ autoregressively based on the hidden states and previously generated tokens. The training objective of the Seq2Seq model is to minimize the negative log-likelihood loss:

$$\mathcal{L}(\theta) = -\sum_{i=1}^{n} \log p(\hat{y}_i = y_i | X, Y_{<i}, \theta), \tag{2}$$

where $\theta$ is learnable model parameters, $\hat{y}_i$ is the $i$-th token predicted by the model, and $Y_{<i} = y_1, y_2, ..., y_{i-1}$ denotes a set of tokens before the $i$-th token $y_i$.

**Seq2Edit Model.** Due to the overlap between $X$ and $Y$, it is inefficient to perform autoregressive generation on the entire target $Y$, while the Seq2Edit model is a good choice. The Seq2Edit model typically regards GEC as a sequence labeling task, which is composed of a BERT-like encoder and a simple classifier stacked on the top. Firstly, a pre-defined set of tags is needed to represent editing operations. In general, this set of tags includes generic edits (e.g.*$KEEP* for keeping the current token unchanged, *$DELETE* for deleting the current token) and token-dependent edits(e.g. *$APPEND* for appending a new token next to the current token, *$REPLACE* for replacing the current token with another token). Considering that the vocabulary of tags linearly increases with token-dependent edits, it is common to set a moderate tag vocabulary based on the frequency of editing to balance the coverage of editing and model size. Then, the initial sentence pairs are converted into sentence-tag pairs of identical length. Specifically, align the target sentence $Y$ with the source sentence $X$ by minimizing the modified Levenshtein

CCL 2024

distance, and then convert it into a sequence of tags $T = t_1, t_2, ..., t_m$. In training, the general objective function of the Seq2Edit methods is to minimize the negative log-likelihood loss for the tag sequence:

$$\mathcal{L}(\theta) = -\sum_{i=1}^{n} \log p(\hat{t}_i = t_i | X, \theta), \tag{3}$$

where $\hat{t}_i$ is the $i$-th tag predicted by the model. During the inference stage, Seq2Edit methods predict a tagging sequence $T$, then perform editing operations on the source sentence $X$ via post-processing to obtain the predicted output $Y$.

### 3.3 Iterative Optimization

Since traditional GEC models are sensitive to some high-error types with high-probability distributions but ignore low-probability ones, we first employ LLMs to analyze grammar errors that these models did not rectify. Then, we exploit LLMs to generate sentences containing uncorrected errors based on the analysis results. Finally, the newly generated corpus is used to iteratively re-train traditional GEC models, thus further enhancing their accuracy.

In the iterative optimization process, we incorporate two different data sources: the pseudo corpus $\mathcal{D}_{\mathtt{pse}}$, and a small human-annotated corpus known as golden data, represented as $\mathcal{D}_{\mathtt{ann}}$. To bridge the distribution gap between the pseudo corpus and the human-annotated corpus, we make use of $\mathcal{D}_{\mathtt{ann}}^{\mathtt{dev}}$. Subsequently, $\mathcal{D}_{\mathtt{ann}}^{\mathtt{test}}$ is employed to evaluate the efficiency of GEC models.

**Algorithm 1** An algorithm with Iterative Optimization

**Input:** Pseudo corpus $\mathcal{D}_{\mathtt{pse}}$; Development dataset $\mathcal{D}_{\mathtt{ann}}^{\mathtt{dev}}$; GEC models $\mathtt{f}$;
The LLM model $\mathbb{P}_{\mathtt{LLM}}$; The number of iterative optimization rounds $\mathtt{N}$;
Iterative optimization prompts $\mathcal{T}_{\mathtt{IE}}$.
**Output:** Re-training corpus $\mathcal{D}_{\mathtt{retrain}}$; Optimized GEC models $\mathtt{f}'$.

$\mathcal{D}_{\mathtt{add}}^{0} \leftarrow \emptyset$
**for** `r in range(N)` **do**
  `Init(f);` ▷ Initialize GEC models with annotated training corpus
  $\mathcal{D}_{\mathtt{retrain}}^{\mathtt{r}} \leftarrow \mathcal{D}_{\mathtt{pse}} \cup (\cup_{\mathtt{i=1}}^{\mathtt{r}} \mathcal{D}_{\mathtt{add}}^{\mathtt{i}})$
  `Retraining(f,`$\mathcal{D}_{\mathtt{retrain}}^{\mathtt{r}}$`)`
  $\mathcal{D}_{\mathtt{pre}}^{\mathtt{r}} \leftarrow \mathtt{Correct}\{\mathtt{f}'(\mathcal{D}_{\mathtt{ann}}^{\mathtt{dev}} \mid \mathcal{D}_{\mathtt{retrain}}^{\mathtt{r}})\}$ ▷ Evaluate GEC models with synthetic corpus
  $\mathcal{D}_{\mathtt{add}}^{\mathtt{r+1}} \leftarrow \emptyset$
  **for** `each (x,y)` $\in \mathcal{D}_{\mathtt{pre}}^{\mathtt{r}}$ **do**
    $(\mathtt{x}_{\mathtt{add}}, \mathtt{y}_{\mathtt{add}}) \sim \mathbb{P}_{\mathtt{LLM}}(\cdot | \mathcal{T}_{\mathtt{IE}}(\mathtt{x}_{\mathtt{pre}}, \mathtt{y}_{\mathtt{ann}}))$ ▷ Identify errors and generate coupus
    $\mathcal{D}_{\mathtt{add}}^{\mathtt{r+1}} \leftarrow \mathcal{D}_{\mathtt{add}}^{\mathtt{r+1}} \cup \{(\mathtt{x}_{\mathtt{add}}, \mathtt{y}_{\mathtt{add}})\}$
  **end for**
**end for**
$\mathcal{D}_{\mathtt{retrain}} \leftarrow \mathcal{D}_{\mathtt{pse}} \cup (\cup_{\mathtt{i=1}}^{\mathtt{N}} \mathcal{D}_{\mathtt{add}}^{\mathtt{i}})$ ▷ Iteration completion and hybridize corpus
$\mathtt{f}' = \mathtt{Re\text{-}training(f,}\mathcal{D}_{\mathtt{retrain}}\mathtt{)}$

We present the whole process of iterative optimization in Algorithm 1. First, we set the additional corpus $\mathcal{D}_{\mathtt{add}}^{0}$ to an empty set in the 0-th round. The iterative optimization process starts in the second line, represented by a for-loop, where each iteration of the for-loop corresponds to one optimization round. In our experiment, we conduct 4 rounds of iterative optimization tests. For each loop, the parameters of GEC models are initialized with an annotated training corpus. Second, we combine the pseudo corpus $\mathcal{D}_{\mathtt{pse}}$ and the additional corpus $\mathcal{D}_{\mathtt{add}}^{\mathtt{r}}$ from the previous $r$-th rounds to form the re-training corpus $\mathcal{D}_{\mathtt{retrain}}^{\mathtt{r}}$ in the $r$-th round. Then, all GEC model parameters are updated by the re-training corpus $\mathcal{D}_{\mathtt{retrain}}^{\mathtt{r}}$. Third, we evaluate optimized GEC models $\mathtt{f}'$ on the development dataset $\mathcal{D}_{\mathtt{ann}}^{\mathtt{dev}}$ and combine the predicted sentences $\mathtt{x}_{\mathtt{pre}}$ of GEC models to form the set of predicted sentences $\mathcal{D}_{\mathtt{pre}}^{\mathtt{r}}$ in the $r$-th round. We set the additional corpus $\mathcal{D}_{\mathtt{add}}^{\mathtt{r+1}}$ in the `r+1`-th round to an empty set. Fourth, we utilize $\mathcal{T}_{\mathtt{IE}}(\mathtt{x}_{\mathtt{pre}}, \mathtt{y}_{\mathtt{ann}})$ prompt to motivate the LLMs to analyze whether there are still grammatical errors in

CCL 2024

the predicted sentences $x_{pre}$ through optimized GEC models. Based on the analysis results, we employ the LLMs generate sentence paris $(x_{add}, y_{add})$ resembling the predicted sentence $x_{pre}$ with the annotated sentence $y_{ann}$. Finally, We combine the pair of generated sentence $(x_{add}, y_{add})$ to form the additional corpus $\mathcal{D}_{add}^{r+1}$ of the $r+1$ round. After $N$ iterative optimizations, we hybridize the generated additional corpus $\mathcal{D}_{add}^{N}$ with the pseudo corpus $\mathcal{D}_{pse}$ to form the ultimate re-training corpus $\mathcal{D}_{retrain}$. Then, we retrain GEC models based on the ultimate re-training corpus $\mathcal{D}_{retrain}$ to obtain the final iteratively optimized GEC models. The primary steps of the iterative optimization algorithm involve retraining the GEC models using the current synthesized corpus (line 5) and employing the LLM to extend the predicted sentences to generate additional training corpus (lines 8-11). In summary, the iterative optimization process utilizes LLMs to identify errors within the GEC models and generate a corpus based on this error information, thus helping GEC models find better convergence points.

## 4 Experiments

### 4.1 Experimental Settings

**Datasets.** We adopt the wudao corpus[0] as the unlabeled corpus to generate pseudo corpus for the large language model-guided optimization training method. We employ a combination of the Chinese Lang8 (Zhao et al., 2018a), the HSK (Zhang, 2009), and our generated pseudo corpus by LLMs for traditional model training. MuCGEC-*dev* (Zhang et al., 2022a) is used for hyper-parameter tuning and checkpoint selection. Table1 presents detailed statistical information for the mentioned datasets.

| Dataset | #Sent | #Error | Usage |
|---|---|---|---|
| **Lang8** | 1,220,906 | 1,092,285 (89.5%) | Training |
| **HSK** | 156,870 | 95,320 (60.8%) | Training |
| **WuDao** | 200,000 | - | Training |
| **MuCGEC-*dev*** | 1,125 | 1,069 (95.1%) | Validation |
| **MuCGEC-*test*** | 5,938 | 5,475 (92.2%) | Testing |
| **NLPCC2018-*test*** | 2,000 | 1,983 (99.2%) | Testing |

Table 1: Statistics of the used CGEC datasets. **#Sent** denotes the number of the sentences and **#Error** denotes the number (the percentage) of the erroneous sentences.

**Evaluation Metrics.** For the evaluation metrics, we follow previous work and report word-level performance on the NLPCC2018-*test* using the official MaxMatch scorer and the PKUNLP word segmentation tool provided by Zhao et al. (2018b). For MuCGCE-*test* (Zhang et al., 2022a), we report the character-level scores using the ChERRANT tools.

**Basic Model.** The mainstream CGEC models are primarily categorized into two types, i.e., Seq2Seq and Seq2Edit. We have chosen the Seq2Seq and Seq2Edit models as the baseline representatives. For the Seq2Seq model, we opt for Chinese BART-large (Shao et al., 2021) to generate text, as it can be straightforwardly trained for CGEC. For the Seq2Edit model, GECToR-Chinese (Omelianchuk et al., 2020) stands out as the predominant method employed for CGEC. Building upon the foundation laid by Zhang et al. (2022a), our selection for the encoder in the Seq2Edit method is Chinese-Struct-Bert-Large (Wang et al., 2020). We employ Qwen-14B-Chat (Bai et al., 2023)[1] for pseudo corpus generation and Qwen-72B-Chat-Int4 (Bai et al., 2023)[2] for iterative optimization.

**Implementation Details.** We utilize Chinese BART-large (Shao et al., 2021) and Chinese-Struct-Bert-Large (Wang et al., 2020) to enhance traditional GEC models. For the LLMs inference process, we follow the vllm-project (Kwon et al., 2023)[3] on 4 GeForce RTX 3090 24GB GPUs. Appendix A

[0]https://data.baai.ac.cn/details/WuDaoCorporaText
[1]https://huggingface.co/Qwen/Qwen-14B-Chat
[2]https://huggingface.co/Qwen/Qwen-72B-Chat-Int4
[3]https://github.com/vllm-project/vllm

shows additional implementation specifics and hyperparameter selection. Appendix B gives detailed illustrations about particular prompts utilized in our approach.

### 4.2 Main Results

| System | Extra Data Size | NLPCC2018-*test* P | NLPCC2018-*test* R | NLPCC2018-*test* $F_{0.5}$ | MuCGEC-*test* P | MuCGEC-*test* R | MuCGEC-*test* $F_{0.5}$ |
|---|---|---|---|---|---|---|---|
| **Seq2Edit** (Zhang et al., 2022a) | - | 42.88 | 30.19 | 39.55 | 44.11 | 27.18 | 39.22 |
| **Seq2Seq** (Zhang et al., 2022a) | - | 41.44 | 32.89 | 39.39 | 43.81 | 28.56 | 39.58 |
| **SynGEC** (Zhang et al., 2022b) | - | 49.96 | 33.04 | 45.32 | 54.69 | 29.10 | 46.51 |
| **TemplateGEC** (Li et al., 2023a) | - | 54.50 | 27.40 | 45.50 | - | - | - |
| **Qwen-14B-Chat** (Bai et al., 2023) | - | 24.50 | 41.71 | 26.70 | 22.47 | 34.85 | 24.19 |
| **Qwen-72B-Chat-Int4** (Bai et al., 2023) | - | 27.94 | 33.11 | 28.84 | 25.62 | 30.41 | 26.45 |
| **Direct Noise** (Ye et al., 2023) | 800w | 49.57 | 31.80 | 44.59 | 54.93 | 29.61 | 46.91 |
| **Backtranslation** (Ye et al., 2023) | 800w | 47.64 | 36.43 | 44.88 | 54.82 | 30.27 | 47.17 |
| **Our Seq2Edit** | - | 42.87 | 30.25 | 39.57 | 43.97 | 27.45 | 39.25 |
| ***w/* Iterative Optimization** | 20w | 43.74 | 36.61 | **42.10** | 44.34 | 32.60 | **41.36** |
| **Our Seq2Seq** | - | 49.68 | 31.79 | 44.65 | 54.17 | 28.74 | 46.03 |
| ***w/* Iterative Optimization** | 20w | 52.17 | 31.51 | **46.12** | 55.74 | 29.56 | **47.35** |

Table 2: Overall results on NLPCC2018-*test* and MuCGEC-*test*. The results of Seq2Edit (Zhang et al., 2022a), Seq2Seq (Zhang et al., 2022a), SynGEC (Zhang et al., 2022b), TemplateGEC (Li et al., 2023a), Direct Noise and Backtranslation (Ye et al., 2023) on NLPCC2018-*test* and MuCGEC-*test* are cited from the original papers and other results including Qwen-14B-Chat and Qwen-72B-Chat (Bai et al., 2023) are implemented by us. The best results of two different traditional models are highlighted in bold.

Table 2 shows the main results of NLPCC2018-*test* and MuCGEC-*test* and makes a detailed comparison with previous works. First, we can see that our Seq2Seq and Seq2Edit models have achieved comparable or even better performances than previous works (Zhang et al., 2022a), which verifies the strong ability of our baseline models. Second, compared with the top block, we find that previous methods frequently integrate supplementary information to boost the efficacy of GEC models, including syntactic and confusion set knowledge (Zhang et al., 2022a; Li et al., 2023a). Due to LLMs acquiring a strong understanding of grammar during training, our approach focuses on effectively excavating this knowledge from LLMs and utilizing it to enhance the performance of GEC models. Our approach consistently surpasses all comparison methods in terms of $F_{0.5}$. This result indicates that our method can effectively excavate the grammar knowledge from LLMs and utilize it in Seq2Seq and Seq2Edit models. Third, we find that our large language model-guided iterative optimization training approach leads to 1.47/1.32 improvements in the Seq2Seq model and 2.53/2.11 improvements in the Seq2Edit model compared to the traditional training method. Compared with direct noise and backtranslation augmentation methods, our model can achieve better results with only a smaller pseudo corpus. This highlights the efficacy of our iterative optimization strategy and the small-scale corpus construction method we introduced. Finally, we can observe that our approach performs superior results with Seq2Seq and Seq2Edit models compared to utilizing LLMs directly for grammatical error correction, indicating that employing constraints like error types and confusion sets can guide LLMs to construct corpus effectively. Moreover, we can effectively utilize the strong analytical capabilities of LLMs to enhance the performance of Seq2Seq and Seq2Edit models.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1366-1380, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

### 4.3 Analyses and Discussion

#### 4.3.1 Ablation Study

The main results of our approach stem from three kinds of constraint information, including error types, confusion set, and iterative optimization from LLMs. Hence, it is imperative for us to perform ablation studies on these three types of grammar information to assess their impacts on our method. Table 3 shows a thorough ablation study on the NLPCC2018-*test* and MuCGEC-*test* datasets.

| Method | NLPCC2018-*test* | | | MuCGEC-*test* | | |
|---|---|---|---|---|---|---|
| | **P** | **R** | $F_{0.5}$ | **P** | **R** | $F_{0.5}$ |
| **Seq2Edit** | 43.74 | **36.61** | **42.10** | 44.34 | **32.60** | **41.36** |
| **w/o Error Types** | 43.52 | 35.32 | 41.59 | 43.97 | 32.06 | 40.93 |
| **w/o Confusion Set** | **43.92** | 35.11 | 41.82 | **44.47** | 31.39 | 41.05 |
| **w/o Iterative Optimization** | 42.87 | 36.10 | 41.32 | 43.74 | 32.12 | 40.79 |
| **Seq2Seq** | 51.82 | 32.03 | **46.12** | 55.74 | 29.56 | **47.35** |
| **w/o Error Types** | 51.41 | 31.40 | 45.60 | 55.23 | 29.33 | 46.94 |
| **w/o Confusion Set** | **51.96** | 31.30 | 45.90 | **55.83** | 28.86 | 47.04 |
| **w/o Iterative Optimization** | 49.87 | **33.75** | 45.52 | 54.42 | **30.06** | 46.83 |

Table 3: Results of ablation study on NLPCC2018-*test* and MuCGEC-*test*

First, we can observe that incorporating each type of constraint information leads to notable enhancements in both Seq2Seq and Seq2Edit models when implemented separately, affirming the validity of our decision to acquire data from LLMs. Due to the high proportion of substituted errors in the HSK corpus, the enhancement in the performance of the GEC model does not vary significantly based on error types and constraints related to confusion set information. The primary cause of this result is that the model becomes trapped in spurious patterns during the training process, leading to minimal performance enhancements. Therefore, iterative optimization has brought the greatest improvement for Seq2Seq and Seq2Edit models, indicating that LLMs based on corrected sentences and grammatically correct sentences can generate valuable reference corpora. These outcomes offer valuable insights for Seq2Seq and Seq2Edit models, resulting in substantial benefits for these models. Moreover, the combination of different constraint information leads to sustaining improvements in all GEC models, demonstrating the compatibility between the three types of our designed information.

#### 4.3.2 Distribution on four error types

To further verify the effectiveness of our method in narrowing the gap in error type distribution compared to the golden corpus, we utilize the HSK corpus as the foundation corpus. The error type distribution is illustrated across the four CGEC error categories: missing (M), redundant (R), substitution (S), and word-order (W). We utilize the ChERRANT tool for converting the corpus into the M2 file format and extracting the distribution of error types within the corpus.

The results depicted in Figure 3 demonstrate that our method further narrows the error type distribution gap between the pseudo corpus and the realistic corpus. The constraint information in the confusion set primarily focuses on addressing substituted errors. Therefore, we only replace the part of the corpus that consists of substitution errors generated by error type constraint information. Through iterative optimization information, in cases where the proportion of missing errors is relatively low, we iteratively generate more corpus containing missing errors to further narrow the error type distribution gap between the generated corpora and the realistic corpora. In brief, our method closely aligns with the error type distribution found in the HSK corpus by utilizing constraints that take into account error types, confusion sets, and iterative optimization.

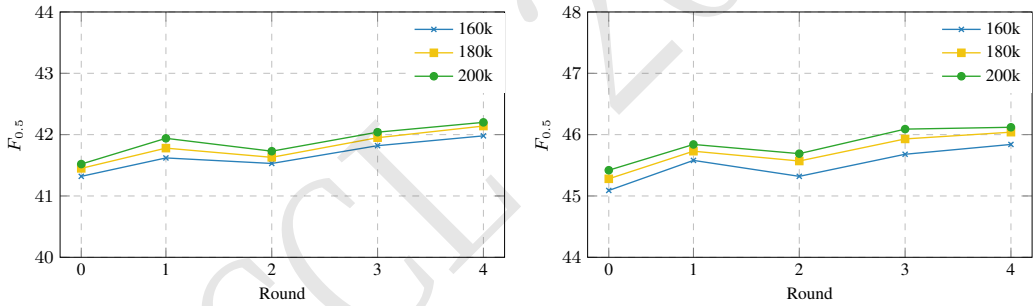
CCL 2024




Figure 3: The results of distribution on four error types

### 4.3.3 Effectiveness of Iterative Optimization

To further investigate the effect of our iterative optimization method on traditional GEC models, we construct three different scales of pseudo corpus (160k, 180k, and 200k sentences) based on error types and confusion sets constraint information, and integrate them into the Lang8 and HSK corpora. Then, we train the GEC model on these three different scales of corpus and perform four rounds of iterative optimization on the NLPCC2018 test set. In other words, we want to use the powerful analytical capabilities of LLMs to analyze whether there are any remaining uncorrected grammatical errors in sentences predicted by the GEC model. Based on the analysis results, we aim to generate a high-quality corpus that is instructive.




(a) Seq2Edit (b) Seq2Seq

Figure 4: Results of our method on NLPCC2018 test set with different pseudo corpus scales.

By observing the results in Figure 4, we find that when grammatically correct sentences are added to LLMs to generate re-training data, it helps to improve the performance of the GEC model. As the number of iterations and corpus size grows, the performance of the GEC model continues to improve. This also indicates that enabling LLMs to view golden sentences during the generation of pseudo corpus enables them to capture the grammatical information from incorrect sentences to correct ones, thereby reducing the probability of GEC model falling into a pattern of illusory modifications. Therefore, we can also understand why there is a significant performance improvement when golden sentences are added to LLM inputs during the generation of training data. In this scenario, LLMs generate sentences in the training data similar to golden sentences in the test set.

## 5 Conclusion

In this paper, we propose a large language model-guided optimization training method to exploit the strong generation and inference capability of LLMs to enhance the performance of traditional GEC

models. On the one hand, we propose using error types and confusion sets as constraint information to guide LLMs in generating diverse pseudo corpus, thus enhancing the robustness of traditional GEC models. On the other hand, we leverage LLMs to infer the predictions of our GEC models to reduce the gap between the error distributions of re-training and realistic corpora. iteratively optimizing our CGEC models. Experimental results on two CGEC datasets demonstrate the effectiveness of our approach on Seq2Edit and Seq2Seq models. Further detailed analysis shows the effectiveness of our proposed method in reducing the distribution of error types, as well as the key role of iterative optimization in enhancing performance. In future research, we will further explore how to fully utilize the grammatical information of large models to enhance the performance of Seq2Seq and Seq2Edit models.

## Acknowledgements

This work was supported by the National Natural Science Foundation of China (U21B2027, 62366027, 62266028, 62306129), Yunnan Provincial Major Science and Technology Special Plan Projects (202103AA080015, 202202AD080003, 202203AA080004), Yunnan Fundamental Research Projects (202401BC070021, 202301AS070047, 202401CF070121), Kunming University of Science and Technology's "Double First-rate" Construction Joint Project (202201BE070001-021, 202301BE070001-027), Yunnan High and New Technology Industry Project (201606). The authors would like to thank the anonymous reviewers for their valuable comments and suggestions for improving this paper.

## References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *CoRR*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *CoRR*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proc. of NeurIPS*.

Christopher Bryant, Zheng Yuan, Muhammad Reza Qorib, Hannan Cao, Hwee Tou Ng, and Ted Briscoe. 2022. Grammatical error correction: A survey of the state of the art. *CoRR*.

Yo Joong Choe, Jiyeon Ham, Kyubyong Park, and Yeoil Yoon. 2019. A neural grammatical error correction system built on better pre-training and sequential transfer learning. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 213–227.

Chenhe Dong, Yinghui Li, Haifan Gong, Miaoxin Chen, Junxin Li, Ying Shen, and Min Yang. 2023. A survey of natural language generation. *ACM Comput. Surv.*, pages 173:1–173:38.

Yaxin Fan, Feng Jiang, Peifeng Li, and Haizhou Li. 2023. Grammargpt: Exploring open-source llms for native chinese grammatical error correction with supervised fine-tuning. In *Proc. of NLPCC*, pages 69–80.

Tao Fang, Shu Yang, Kaixin Lan, Derek F. Wong, Jinpeng Hu, Lidia S. Chao, and Yue Zhang. 2023. Is chatgpt a highly fluent grammatical error correction system? A comprehensive evaluation. *CoRR*.

Masahiro Kaneko and Naoaki Okazaki. 2023a. Controlled generation with prompt insertion for natural language explanations in grammatical error correction. *CoRR*.

Masahiro Kaneko and Naoaki Okazaki. 2023b. Reducing sequence length by predicting edit spans with large language models. In *Proc. of EMNLP*, pages 10017–10029.

Masahiro Kaneko, Masato Mita, Shun Kiyono, Jun Suzuki, and Kentaro Inui. 2020. Encoder-decoder models can benefit from pre-trained masked language models in grammatical error correction. In *Proc. of ACL*, pages 4248–4254.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *CoRR*.

Shota Koyama, Hiroya Takamura, and Naoaki Okazaki. 2021. Various errors improve neural grammatical error correction. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation, PACLIC 2021, Shanghai International Studies University, Shanghai, China, 5-7 November 2021*, pages 251–261.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Jiquan Li, Junliang Guo, Yongxin Zhu, Xin Sheng, Deqiang Jiang, Bo Ren, and Linli Xu. 2022. Sequence-to-action: Grammatical error correction with action guided sequence generation. In *Proc. of AAAI*, pages 10974–10982.

Yinghao Li, Xuebo Liu, Shuo Wang, Peiyuan Gong, Derek F. Wong, Yang Gao, Heyan Huang, and Min Zhang. 2023a. TemplateGEC: Improving grammatical error correction with detection template. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6878–6892.

Yinghui Li, Haojing Huang, Shirong Ma, Yong Jiang, Yangning Li, Feng Zhou, Hai-Tao Zheng, and Qingyu Zhou. 2023b. On the (in)effectiveness of large language models for chinese text correction. *CoRR*.

Yinghui Li, Zishan Xu, Shaoshen Chen, Haojing Huang, Yangning Li, Yong Jiang, Zhongli Li, Qingyu Zhou, Hai-Tao Zheng, and Ying Shen. 2023c. Towards real-world writing assistance: A chinese character checking benchmark with faked and misspelled characters. *CoRR*.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem N. Chernodub, and Oleksandr Skurzhanskyi. 2020. Gector - grammatical error correction: Tag, not rewrite. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications, BEA@ACL 2020, Online, July 10, 2020*, pages 163–170.

Fanyi Qu and Yunfang Wu. 2023. Evaluating the capability of large-scale language models on chinese grammatical error correction task. *CoRR*.

Yunfan Shao, Zhichao Geng, Yitao Liu, Junqi Dai, Fei Yang, Li Zhe, Hujun Bao, and Xipeng Qiu. 2021. CPT: A pre-trained unbalanced transformer for both chinese language understanding and generation. *CoRR*.

Yixiao Song, Kalpesh Krishna, Rajesh Bhatt, Kevin Gimpel, and Mohit Iyyer. 2023. Gee! grammar error explanation with large language models. *CoRR*.

Felix Stahlberg and Shankar Kumar. 2021. Synthetic data generation for grammatical error correction with tagged corruption models. In *Proc. of EACL*, pages 37–47.

Wei Wang, Bin Bi, Ming Yan, Chen Wu, Jiangnan Xia, Zuyi Bao, Liwei Peng, and Luo Si. 2020. Structbert: Incorporating language structures into pre-training for deep language understanding. In *Proc. of ICLR*.

Yu Wang, Yuelin Wang, Kai Dang, Jie Liu, and Zhuo Liu. 2021. A comprehensive survey of grammatical error correction. *ACM Trans. Intell. Syst. Technol.*, pages 65:1–65:51.

Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, et al. 2023a. Instructuie: Multi-task instruction tuning for unified information extraction. *arXiv preprint arXiv:2304.08085*.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023b. Self-instruct: Aligning language models with self-generated instructions. In *Proc. of ACL*, pages 13484–13508.

Haoran Wu, Wenxuan Wang, Yuxuan Wan, Wenxiang Jiao, and Michael R. Lyu. 2023. Chatgpt or grammarly? evaluating chatgpt on grammatical error correction benchmark. *CoRR*.

Jingheng Ye, Yinghui Li, Yangning Li, and Hai-Tao Zheng. 2023. Mixedit: Revisiting data augmentation and beyond for grammatical error correction. In *Proc. of EMNLP Findings*, pages 10161–10175.

CCL 2024

Yue Zhang, Zhenghua Li, Zuyi Bao, Jiacheng Li, Bo Zhang, Chen Li, Fei Huang, and Min Zhang. 2022a. MuCGEC: a multi-reference multi-source evaluation dataset for Chinese grammatical error correction. In *Proc. of NAACL*, pages 3118–3130.

Yue Zhang, Bo Zhang, Zhenghua Li, Zuyi Bao, Chen Li, and Min Zhang. 2022b. Syngec: Syntax-enhanced grammatical error correction with a tailored gec-oriented parser. In *Proc. of EMNLP*, pages 2518–2531.

Yue Zhang, Leyang Cui, Deng Cai, Xinting Huang, Tao Fang, and Wei Bi. 2023. Multi-task instruction tuning of llama for specific scenarios: A preliminary study on writing assistance. *CoRR*.

Baolin Zhang. 2009. Features and functions of the hsk dynamic composition corpus. *International Chinese Language Education*, pages 71–79.

Yuanyuan Zhao, Nan Jiang, Weiwei Sun, and Xiaojun Wan. 2018a. Overview of the NLPCC 2018 shared task: Grammatical error correction. In *Natural Language Processing and Chinese Computing - 7th CCF International Conference, NLPCC 2018, Hohhot, China, August 26-30, 2018, Proceedings, Part II*, pages 439–445.

Yuanyuan Zhao, Nan Jiang, Weiwei Sun, and Xiaojun Wan. 2018b. Overview of the NLPCC 2018 shared task: Grammatical error correction. In *Proc. of NLPCC*, pages 439–445.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. *CoRR*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023a. LIMA: less is more for alignment. In *Proc. of NeurIPS*.

Houquan Zhou, Yumeng Liu, Zhenghua Li, Min Zhang, Bo Zhang, Chen Li, Ji Zhang, and Fei Huang. 2023b. Improving seq2seq grammatical error correction via decoding interventions. In *Proc. of EMNLP Findings*, pages 7393–7405.

## A Implementation Details and Hyperparameters

For Seq2Edit model and Seq2Seq model training, following Zhang et al. (2022a), we list the main hyper-parameters in Table 4 and Table 5. To determine the optimal balancing weights for the training objective, we experiment with various values for $\alpha$ within 0.5, 1.0, 1.2, 2.0 and $\beta$ within 0.5, 1.0, 2.0. Our experiments reveal that the configuration with $\alpha$ = 1.0 and $\beta$ = 1.0 achieves the highest $F_0.5$ score on NLPCC2018-*test* and MuCGEC-*test*, which is used as the default settings in all our experiments.

| Configurations | Values |
|---|---|
| Backbone | Chinese-StructBERT |
| Devices | 1 GeForce RTX 3090 |
| Number of max epochs | 20 |
| Number of cold epochs | 2 |
| Optimizer | Adam |
| Cold learning rate | $1\times10^{-3}$ |
| Learning rate | $1\times10^{-5}$ |
| Batch size | 128 |
| Patience | 3 |

Table 4: Hyper-parameter values of Seq2Edit.

| Configurations | Values |
|---|---|
| Backbone | Chinese-BART-large |
| Devices | 4 GeForce RTX 3090 |
| Epochs | 20 |
| Optimizer | Adam |
| Learning rate | $3\times10^{-5}$ |
| Batch size per GPU | 4096 tokens |
| Warmup updates | 2000 |
| Max source length | 1024 |
| Dropout | 0.2 |
| Dropout-src | 0.2 |
| Beam size | 12 |

Table 5: Hyper-parameter values of Seq2Seq.

## B Designed Prompts

In order to guide LLMs to generate high-quality corpus under the constraints of error types and confusion set, we have carefully designed instruction prompts based on the characteristics of constructing the CGEC task corpus. The prompt for generation is shown in Table 6 and the prompt for iterative optimization is shown in Table 7.

We mainly cover missing, redundant, substitution, and word-order as the primary types of errors. For missing errors, we prompt the LLMs to delete one or more words in the input sentence. For redundant errors, we prompt the LLMs to repeat one or more words in the input sentence, but only from the words present in the input sentence. For substitution errors, we prompt the LLMs to replace one or more of the words in the input sentence with synonyms, homophones, or words that are similar in form to that word. For word-order errors, we prompt the LLMs to rearrange the order of words in the input sentence, resulting in a disorganized sentence structure. It is important to keep the words in the sentence unchanged. The corpus generation methods for the various error types mentioned above correspond to the "operation" section of the prompt for constructing corpus.

| |
|---|
| 你的目标是针对输入句子的标点、词语和句法等，在其中巧妙地引入实用、合理且忠实的中文语法错误。实用性要求构造的句子有助于提升中文语法纠错系统的性能和帮助人们理解正确的中文语法规则。合理性要求构造后的句子能被中文语法纠错系统纠正。忠实性要求构造方式是基于指定的中文错误类型和混淆集信息。<br>指定的中文语法错误类型为[error_types]。请遵循[工作方式]，严格保持句子编号不变。下面让我们一步步地按照[步骤]，针对[输入句子]，生成中文语法错误的句子。<br>[混淆集信息]<br>{con_info}<br>[步骤]<br>Step1: 初步分析输入句子的标点、词语和句法等。<br>Step2: 准确分析出输入句子中每个词语的词性。<br>Step3: 根据分析的词性、中文语法错误类型和混淆集信息，针对输入句子中的词语，{operations}。<br>Step4: 分析一下构造句子中的标点、词语和句法等，确保句子中包含中文语法错误。<br>Step5: 无需进行额外的解释，直接输出构造的句子。<br>[工作方式]<br>[ID][输入句子]<中文语法正确的句子。><br>[ID][输出句子]<构造的句子。><br>请严格遵循你的工作方式。现在开始对输入的句子进行改动：<br>[ID][输入句子]<input_sentence> |

Table 6: Our designed prompt for constructing corpus.

CCL 2024

你的目标是分析纠错模型预测的句子中的标点、词语和句法等，判断其是否含有未纠正的中文语法错误。如果有，生成一对实用、合理且忠实的句子对。如果没有，则回复[纠错完全]。实用性要求构造的句子有助于提升中文语法纠错系统的性能和帮助人们理解正确的中文语法规则。合理性要求构造后的句子能被中文语法纠错系统纠正。忠实性要求基于未纠正的中文语法错误进行构造。
请遵循[工作方式]，严格保持句子编号不变。下面让我们一步步地按照[步骤]根据[预测句子]和[语法正确句子]的内容生成句子对，无需输出解释过程，直接给出结果。
[步骤]
Step1: 初步分析纠错模型预测后句子中的标点、词语和句法等。
Step2: 准确分析出纠错模型预测后句子中每个词语的词性。
Step3: 根据分析的词性和语法正确句子，判断预测句子中是否还含有未纠正的中文语法错误。如果有则执行下面的步骤，否则回复[纠错完全]。
Step4: 生成类似句子对，其包含语法纠错模型未纠正的中文语法错误的句子和正确句子。
Step5: 检查生成的句子对，保证其包含纠错模型未纠正的中文语法错误。
[工作方式]
[ID][预测句子]<被中文语法纠错模型预测的句子。>
[ID][语法正确句子]<对应语法完全正确的句子。>
[ID][构造的错误句子]<包含语法纠错模型未纠正的中文语法错误的句子。>
[ID][构造的正确句子]<中文语法正确的句子。>
请严格遵循你的工作方式。现在开始根据预测句子和语法正确句子生成类似的句子对：
[ID][预测句子]<predicted_sentence>
[ID][语法正确句子]<annocated_sentence>

Table 7: Our designed prompt for iterative optimization.

# Pattern Shifting or Knowledge Losing? A Forgetting Perspective for Understanding the Effect of Instruction Fine-Tuning

**Chunkang Zhang**[1,3]*, **Boxi Cao**[1,3], **Yaojie Lu**[1], **Hongyu Lin**[1]†, **Liu Cao**[4]
**Ke Zeng**[4], **Guanglu Wan**[4], **Xunliang Cai**[4], **Xianpei Han**[1,2], **Le Sun**[1,2]
[1]Chinese Information Processing Laboratory [2]State Key Laboratory of Computer Science
Institute of Software, Chinese Academy of Sciences, Beijing, China
[3]University of Chinese Academy of Sciences, Beijing, China
[4]Meituan, Beijing, China
{zhangchunkang2021,boxi2020,luyaojie,hongyu,xianpei,sunle}@iscas.ac.cn
{liucao,zengke02,wanguanglu,caixunliang}@meituan.com

## Abstract

Instruction Fine-Tuning(IFT) emerges as an essential step of training large language models to robustly carry out tasks of interest. However, there lacks a systematic investigation about the underlying mechanisms of instruction fine-tuning, particularly on the forgetting phenomenon after IFT, known as alignment tax. Therefore, to understand the mechanism of IFT from the forgetting perspective, we investigate the alternation of the text pattern and knowledge within models throughout the entire IFT process. Specifically, we *restore* fine-tuned models to their base version by training them on the data sharing a similar distribution with the pre-training corpus and compare their results Our experiment indicates that there is a stage transition of forgetting during IFT process: (1) **Pseudo Forgetting**: in this stage, models mainly shift their familiar text pattern away from pre-training data format while the world knowledge is preserved. Consequently, models will recover to their original performance when they are restored to the base version. (2) **Actual Forgetting**: in this stage, models forget the acquired knowledge as well. Therefore, they fail to reach the original performance even if they are restored to the base version.

## 1 Introduction

Instruction Fine-Tuning (IFT) has emerged as an indispensable process during the development of Large Language Models (LLMs) (Touvron et al., 2023a; Touvron et al., 2023b; OpenAI et al., 2023). By training LLMs in formatted input-output pairs, IFT enables them to demonstrate extraordinary capabilities on unseen tasks even in zero-shot settings(Wu et al., 2022). Since IFT requires fewer data compared with the pre-training stage and delivers superior outcomes(Wu et al., 2022), it has become an essential approach to enhance LLMs.

However, previous studies(Ouyang et al., 2022; Bai et al., 2022) find that fine-tuned models fail to solve the problems which their base versions could solve, resulting in performance degradation in specific tasks. Such a seemingly catastrophic forgetting(McCloskey and Cohen, 1989) phenomenon after IFT is reported as alignment tax(Ouyang et al., 2022).

Unfortunately, the underlying mechanism of such forgetting remains an open problem and receive distinct opinions. On the one hand, some researchers (Kotha et al., 2023; Jain et al., 2023;Kung and Peng, 2023) indicate that LLMs merely learn the superficial pattern of IFT data and forgetting originates from suppressing LLMs performance in areas out of distribution with the IFT data. On the other hand, other researchers(Burns et al., 2023;Yin et al., 2023;Lin et al., 2023) suggest that IFT enables LLMs to express the knowledge gained during pre-training in the format of instruction-output pairs, where forgetting originates from trading base models‘ generality for such speciality. The absence of such a

*Work was done during the internship at Meituan.
†Corresponding Author.

©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1381-1394, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

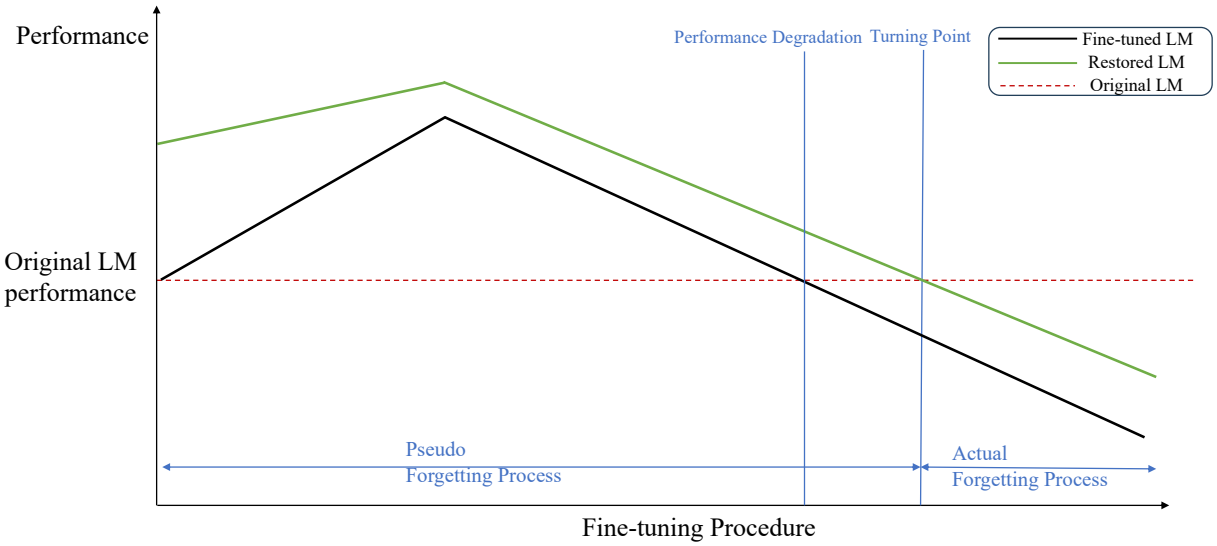



Figure 1: The illustration of theoretical instruction tuning process, during the initial stage, fine-tuned LLM is better than the original. Subsequently, LLM begins to fall behind the original base LLM while restoring it to the original text pattern can bring back the performance. In the late stage, even if LLM is restored to its original pattern, the performance is still worse than the original

systematic and unanimous analysis on the alignment tax limits our potential to develop a better IFT training strategy or thoroughly evaluate LLMs, which consequently delays improvements and further application of LLMs. Therefore, a comprehensive analysis of the underlying core factors behind forgetting after IFT is significant.

To this end, this paper aims to reveal the mechanism of IFT from the forgetting perspective. Specifically, this paper will explore two specific research questions:

- **RQ1: Does the forgetting effect of IFT follow specific patterns?**
- **RQ2: What is the characteristic of these forgetting patterns?**

For RQ1, we propose to restore fine-tuned LLMs to base LLMs in order to investigate whether the world knowledge(Touvron et al., 2023a) is retained in LLMs after IFT. Specifically, we collect data from the corpus base LLMs are pre-trained on, which shares a similar data distribution with the pre-training data. We then sample them to form the *restoring data* for training, which aims in restoring the LLMs familiar text pattern from instruction text to plain text, and *probing data* to measure whether the familiar text pattern is recovered to plain text. Subsequently, we apply several benchmarks to compare the performance between base, IFT and restored LLMs. We find that during the early IFT process, LLMs forget the original text pattern and is more familiar with the text pattern of fine-tuning data, we call such an alternation of text pattern as **pattern shifting**. After restoring the text pattern to the original plain text format, their recovered performance is comparable with base models. Such results indicates that the world knowledge still preserves and the drop originates from the forget of the original text pattern, which we refer to as *Pseudo Forgetting*. As for the late training process, even if the familiar text pattern is restored, the clear gap between recovered models and base models reveals that both text pattern and knowledge are forgotten, which we refer to as *Actual Forgetting*.

As for RQ2, we further investigate the probability distribution of the answer to the benchmark to further reveal the characteristic of forgetting pattern. For the pseudo forgetting stage, we analysis the probability and entropy of the answer across the whole vocabulary and find that the overall desired output generation probability across the whole vocabulary of LLMs begins to converge at high outcome during the early process of pseudo forgetting. We call such phenomenons as *Format Alignment*. After reaching a high generation probability, the entropy of the next generated tokens keeps altering, indicating that models are aligning their preserved knowledge to the desired output format, refered to as *Knowledge*

CCL 2024

*Alignment*. The entropy keeps declining as the training process continues, revealing that models are becoming overconfident, which may cause the actual forgetting. As for the actual forgetting stage, we conduct experiment to measure the memorization of the instruction fine-tuning data and confirm that over memorizing training data is another reason behind actual forgetting.

By summarizing and analyzing results from the performance of different models and instruction data in knowledge benchmark and probing data, we find that there exists a stage of transition in forgetting and shares specific characteristic as followed:

- LLMs are undergoing **Pseudo Forgetting** process during the early stage of IFT. We find that LLMs mainly forget the text pattern of the pre-training data. Specifically, once the fine-tuned LLMs are restored, their performances exceeds the base LMMs.
- LLMs are undergoing **Actual Forgetting** process during the late stage of IFT. We find that LLMs not only forget the text pattern, but also lose their world knowledge. Specifically, we observe a notable gap from base LLMs in performance, even if the performance increase when these LLMs are restored.
- LLMs finish **Format Alignment** early and begins **Knowledge Alignment** during the pseudo forgetting. We conclude that models initially align their output pattern with the questions, which we refer to as format alignment, and continue altering their answers after the completion of format alignment, indicating LLMs are undergoing knowledge alignment.

## 2 Related Work

**Mechanism of Instruction Fine-tuning** Instruction fine-tuning has already become an essential part of developing large language models. However, there is not a uniform opinion towards the mechanism of instruction fine-tuning. On the one hand, some researchers believe that LLMs merely learn the superficial pattern of instruction(Kung and Peng, 2023;Yin et al., 2023;Gupta et al., 2022), On the other hand, other researchers believe that instruction fine-tuning unlocks the potential LLMs acquire during pre-training stage(Burns et al., 2023;Wang et al., 2023). Such disagreement towards the mechanism of IFT restricts the thorough understanding of its theoretical value.

**Distributional Shift** the pre-training and fine-tuning framework has already realized significant success across a variety of applications (Devlin et al., 2019; Radford et al., 2021). Nonetheless, the deployment of pre-trained models into actual real-world settings and their subsequent fine-tuning often leads to a prevalent dilemma: the models encounter novel instances from a target distribution that diverges from the one used during fine-tuning (Andersen and Maalej, 2022; Goyal et al., 2022; Zhang and Ré, 2022). To tackle this problem, a range of solutions have been suggested. For example, Cha et al. (2021b) advocates for the utilization of a weight ensemble that combines the pre-trained and fine-tuned models to improve performance on out-of-distribution (OOD) data. Another method, introduced in (Kumar et al., 2022), is the LP-FT approach, which entails commencing with a pre-trained feature extractor that is paired with a reasonably competent classifier. This initial step is crucial, especially when the classifier begins with a random initialization, as the pre-trained features could quickly become skewed to fit the random classifier during fine-tuning, thereby intensifying the catastrophic forgetting issue.

**Catastrophic Forgetting** Traditional neural network is prone to forgetting the knowledge from a previously learned capabilities upon learning a new task(McCloskey and Cohen, 1989; McClelland et al., 1995; French, 1999). To address such issues, numerous strategies have been developed, such as parameter penalization(Kirkpatrick et al., 2017), continue learning(Cha et al., 2021a; Peng and Risteski, 2022), knowledge distillation(Rebuffi et al., 2017), lifelong learning(Silver et al., 2013;Fischer, 2000) and so on. With the arrival of Large Language Models, research towards catastrophic forgetting mainly focuses on revealing the mechanism of catastrophic forgetting behind continual fine-tuning(Luo et al., 2023;Scialom et al., 2022;Zeng et al., 2023) or continual pre-training(Xia et al., 2023;Tirumala et al., 2022). Therefore, the general forgetting of the pre-train knowledge during instruction tuning remains an open problem,

CCL 2024

Kotha et al. (2023) indicates catastrophic forgetting originates from suppressing LLMs performance on data out of distribution with instruction fine-tuning data while Lin et al. (2023) proposes trade-off between speciality of instruction tuning data and generality of pre-traing is the factor.

## 3 Preliminaries

This section mainly introduces the process of the experiment testbed and corresponding setting to help better understand effect of the finding. Specifically, this section introduces the techniques and details of restoring stage in section 3.1, followed by the introduction of the procedure about testbed in section 3.2. the specific experiment settings are presented in Section 3.3

### 3.1 Restoring the Fine-tuned Model

We propose to restore fine-tuned models to its base versions with data that shares a similar distribution pattern with their pre-training corpus to determine if the knowledge is retained within LLMs. After fine-tuning LLMs on instruction format data, they shift their familiar text distribution patterns to instruction and become less familiar with the pre-training text. Thus the benchmark performance does not thoroughly reflect the related world knowledge within LLMs, it is likely to be related with other phenomenons, such as whether LLMs understand the question of specific task or learn the correct output format. Therefore, We train the fine-tuned LLMs with text resembling their original pre-training data to restore their familiar text pattern from instruction data to the original plain text to ensure the existence of certain knowledge in LLMs.

To be more specific, we control the overall tokens of the restoring training data to be the same as those of IFT data in one epoch, since the main purpose of the restoring stage is to recover the familiar text pattern of the LLMs from instruction to the pre-training data instead of injecting knowledge as the pre-training stage does. Controlling the amount of tokens to be the same as IFT data can reflect whether the restoring process is possible based on the comparison with the base LLMs on probing data as well as a more convinced confirmation of forgetting stages.

In order to further control the distributional shift during restoring stage, we rank the data based on perplexity of llama from low to high and select the top ones as the restoring data. By doing so, the selected data is more possible to be pre-trained before, thus no more new knowledge of text pattern will be introduced, assuring the restoring process just restore the text pattern.

### 3.2 Testbed Procedure

This paper leverages world knowledge probing benchmark as a testbed in order to investigate the forgetting mechanism of LLMs, which focuses on determining whether the world knowledge exists after IFT. As illustrated in Figure **??**, to evaluate how well LLMs preserves their learned knowledge after IFT, we fine-tune these LLMs with instruction data, probe them with knowledge benchmarks and check their answers through out the whole IFT process. To comprehensively determine the existence of specific knowledge, we propose to restore the fine-tuned LLMs to be close to the base version, which controls the effect of pattern shifting of IFT on knowledge probing result.

### 3.3 Experiment Setting

**Language Model** We conduct experiments on transformer-based architecture such as LLaMa2-7b and LLaMa2-13b(Touvron et al., 2023b), since numerous LLMs are derived from these models. We believe findings based on these models will be more general.

**Training Data** The construction of our testbed requires both instruction fine-tuning process and restoring process. For the instruction tuning process, we utilize alpaca(Taori et al., 2023) and vicuna(Zheng et al., 2023) as the instruction training dataset. As for the pattern restoring process, we randomly sample data from Common Crawl-MAIN-2023-40 corpus(Touvron et al., 2023b), which shares a similar distribution with the pre-training corpus that LLaMa2 is pre-trained on to avoid another distribution shifting. Subsequently, we sample about top forty thousand passages which shares the same amount of

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1381-1394, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

tokens as the instruction tuning data with a low perplexity and ten thousand passages as the probing data, indicating whether fine-tuned models are restored to their original pattern.

**Benchmark and Evaluation Metric** We construct our experiment testbed based on ARC(Min, 2023) as well as MMLU(Hendrycks et al., 2021) for probing world knowledge. ARC is designed to evaluate the common sense knowledge while MMLU is a comprehensive benchmark containing variety of domain knowledge. These benchmarks test distinct world knowledge type within LLMs, leading to a more comprehensive finding. As for the evaluation metric, apart from the accuracy of MMLU and ARC, we apply perplexity on formally sampled probing data to measure whether the pattern is restored following the setting of Xia et al. (2023).

**Training Details** During the instruction tuning process, we only calculate loss on outputs, setting epoch to 10, learning rate to $2e^{-5}$ and batch size to 256. As for the restoring process, we set epoch to 1, learning rate to $3e^{-5}$ and batch size to 256. We use FSDP(Zhao et al., 2023) for all the training stage and all experiments are implemented on Nvidia A100-80GB GPUs.

## 4 Stage Transition of Forgetting during Instruction Fine-Tuning

This section proposes results and findings of the forgetting phenomenon. Experiment results lead to following conclusions that forgetting in instruction tuning has two stages, pseudo forgetting and actual forgetting Further analysis based on these conclusions are presented in section 4.2.

### 4.1 Result and Findings

**Finding 1.** *LLMs suffer from pseudo forgetting during the early stage of instruction tuning, where world knowledge is retained in LLMs*

To show this, we firstly investigate whether the restoring process capable of restoring IFT models' famliar text pattern to their pre-training data version. Figure 3a is the trajectory of perplexity on probing data of the ten-epoch fine-tuned model, where we find the perplexity declines as the training process continues and conveys to the original base model perplexity. Such a result indicate that the text pattern is restored to the pre-training data style.

From Figure 3b, we find that LLMs' perplexity against pre-training probing data reveals a continuing upward trend along with the instruction tuning process, indicating that the text pattern LLMs learn during the pre-training process is forgotten. Subsequently, the perplexity of different fine-tuned LLMs have all returned to convergence with the base LLM no matter how large the perplexity is. Both of these results reveal the effectiveness of the restoring process on recovering text pattern.

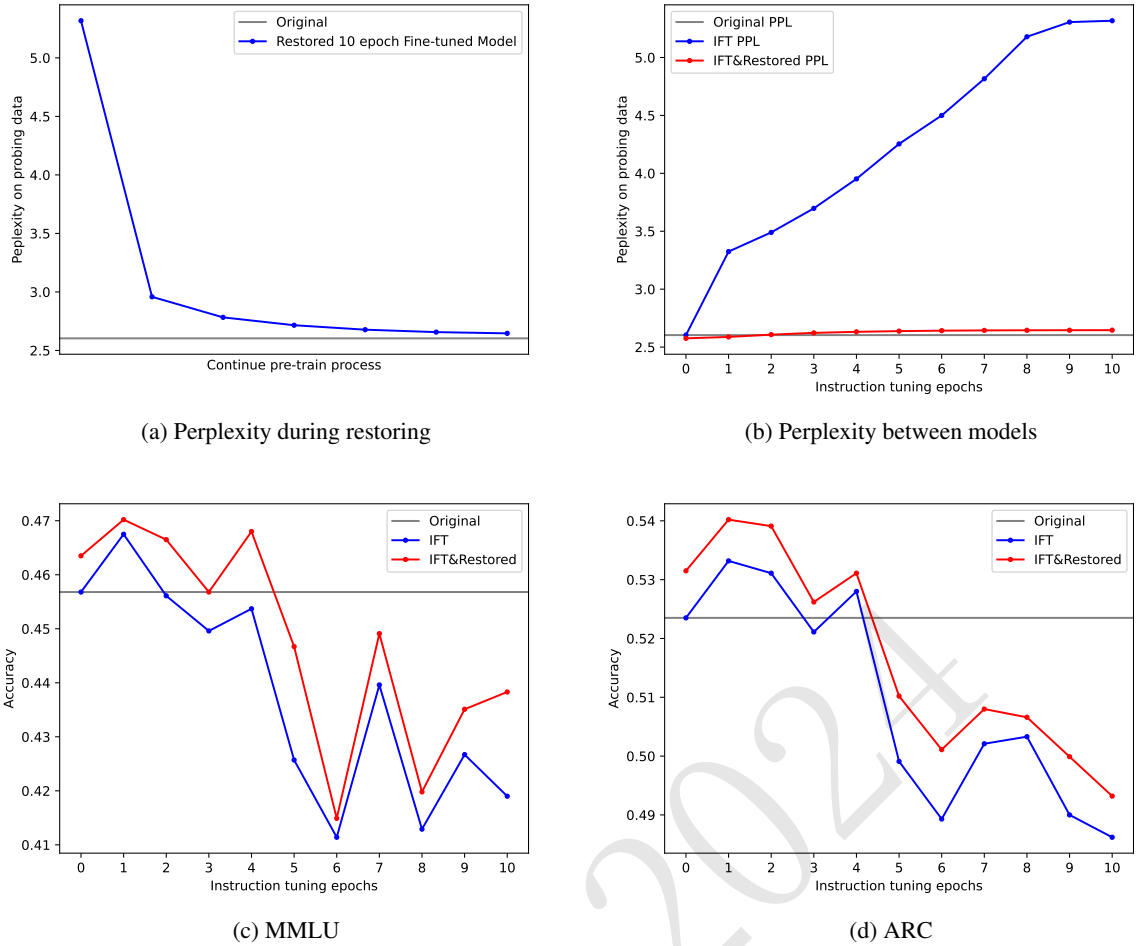
Subsequently, to further determine whether the world knowledge is retained, we conduct experiments on world knowledge benchmark. Figure 3c and Figure 3d demonstrate the performance of IFT model trained on alpaca as well as the restored LLMs, which share almost identical text pattern with base LLMs as revealed in Figure 3b.

From Figure 3c and Figure 3d, we can see that during the initial training process, knowledge is retained in LLMs: 1) despite the majority of the fine-tuned outcomes falling short of the established original baseline, there is a marked enhancement in performance after the restoring stage. 2)In the early training process, the performance of LLMs exceeds original base LLMs when their familiar text patterns have been restored. Both phenomenons indicate that initial degradation after IFT originates from pattern shifting. When the pattern is restored to its original setting, performance is restored near or above the baseline, confirming that the world knowledge within LLMs is retained.

**Finding 2.** *LLMs suffer from actual forgetting during the latter stage of instruction tuning, where LLMs not only shift text pattern away but also lose learned world knowledge*

Figure 3c and Figure 3d illustrate a notable distinction during the late IFT process, indicating the lost of knowledge we refer to as actual forgetting, where performances of LLMs on both benchmarks are notably below the original LLMs even though text patterns are restored. Such phenomenon suggests that IFT induces a more profound and potentially deleterious effect on the models' knowledge base. Together

CCL 2024




(a) Perplexity during restoring

(b) Perplexity between models

(c) MMLU

(d) ARC

Figure 3: Figure 3a illustrates the perplexity on probing data of fine-tuned llama2-7b along with the restoring process, and Figure 3b reveals perpexity on probing data of fine-tuned and restored models. Figure 3c and Figure 3d show the performance of original, sft and restored models on both MMLU and ARC.

these results imply that during the late process of IFT, LLMs begin to forget specific knowledge that were previously acquired during pre-training stage.

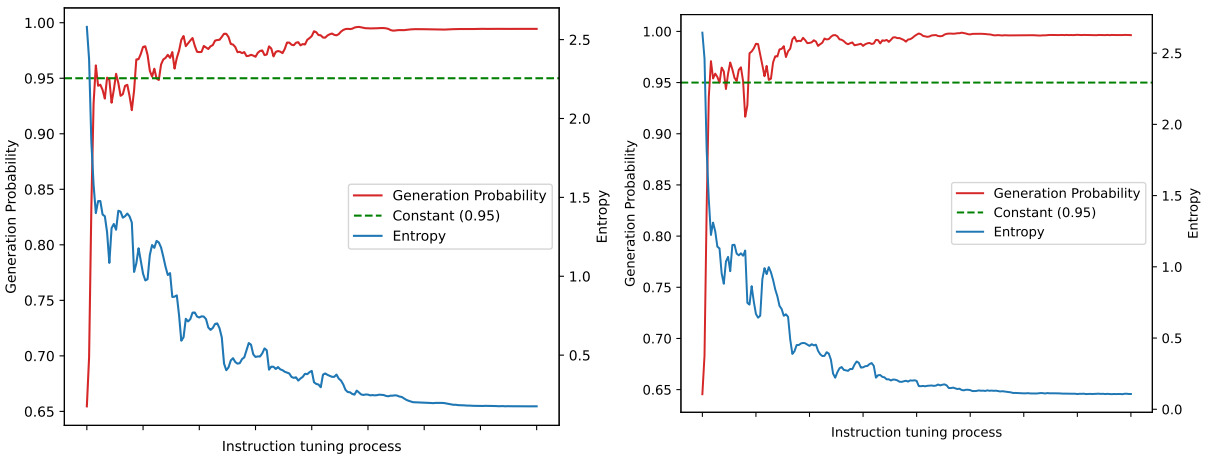
### 4.2 Analysis

During the instruction tuning phase, the empirical evidence suggests that the performance of LLMs experiences an initial surge, surpassing the baseline, before subsequently diminishing and forming a substantial deficit relative to the baseline. Such a performance fluctuation implies that the knowledge embedded within the LLMs is initially more accessible as a result of IFT. However, this activation does not have sustained durability, as indicated by the eventual performance drop. Consequently, when the text pattern of the LLMs is reverted to its original through restoring stage, the performance of the LLMs not only recovers but also exceeds the original baseline, which is possibly the outcome of knowledge activation as Burns et al. (2023) indicates. Thus we move on to investigate the specific characteristic of these forgetting stages.

## 5 Characteristic of Forgetting Stages

In this section, we verify the characteristic of stages during the pseudo forgetting process, we focus more on the trajectory of actual generation probability as well as entropy over the whole vocabulary distribution. We introduce the preliminaries in section 5.1 and the corresponding finding in section 5.2. Our result

CCL 2024




(a) MMLU (b) ARC

Figure 4: Entropy and generation probability across different selection in the vocabulary, where we find the format alignment finish during the pseudo forgetting, while the entropy descends as the training process goes on.

indicates that format alignment completes during the pseudo forgetting stage and begins knowledge alignment. Models begin to be over confident as the training proceeds which might cause the actual forgetting.

### 5.1 Preliminary

**Exact Memorization** Derived from the definition in Tirumala et al. (2022), Let $V$ denote the vocabulary size. Let $C$ denote a set of contexts, which can be thought of as a list of tuples $(s, y)$ where $s$ is an input context (instruction text) and $y$ is the index of the ground truth token in the dataset that completes the block of text. Let $S$ denote the set of input contexts, and let $f : S \rightarrow R^V$ denote a language model. A instruction context $c = (s, y) \in C$ is memorized if $\text{argmax}(f(s)) = y$. We refer to this as exact memorization, thus we define the overall exact memorization score EM as the proportion of $\text{argmax}(f(s)) = y$ through out the whole instruction dataset.

### 5.2 Result and Findings

**Finding 1.** *format alignment is hard to interfere and completes early during pseudo forgetting, while knowledge alignment takes over the pseudo forgetting phase after the completion of format alignment*

To show this, we dive into the generation process of LLMs during IFT. Figure 4a and Figure 4b illustrate the average generation probability of LLMs on benchmark questions as well as the entropy of the possible generation throughout the whole vocabulary.

From Figure 4a and Figure 4b we can see that the generation probability of correct format answer rise up in a few steps and converges. According to the principle of significance in statistics(Yaddanapudi, 2016), p-value less than 0.05 suggests that the model's output was not due to random chance. As depicted in Figure 4a and Figure 4b, models exhibit format alignment early in the training phase, maintaining a probability of generating the anticipated response above 95%. This high probability persists throughout the entire training process demonstrating the model's format alignment is not readily susceptible to disruption.

To further confirm the knowledge alignment, we can conclude from Figure 4a and Figure 4b that after the completion of format alignment, LLMs' internal entropy undergoes a period of fluctuation. This observation suggests a dynamic alteration in the responses provided by the model throughout the training phase. Despite these variations, the probability of the model producing correctly formatted responses remains stable. Such a fluctuating trend means that the distribution of the answer is changing, proving the existence of the model's knowledge alignment.

CCL 2024

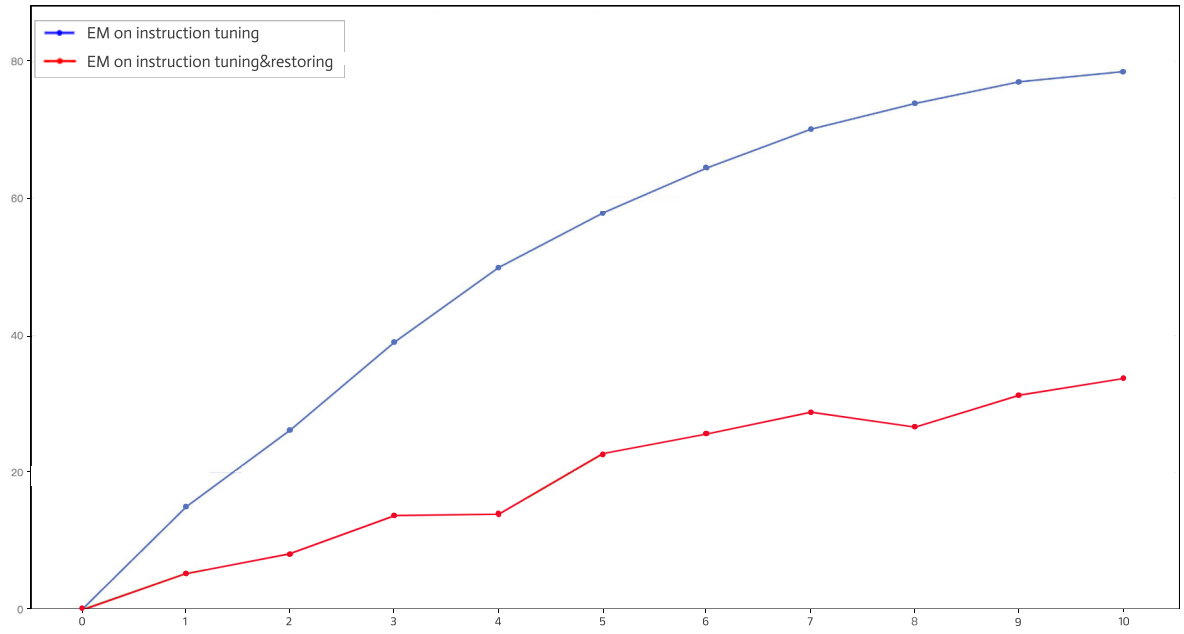



Figure 5: The illustration of alpaca EM along with the training epochs

**Finding 2.** *LLMs are becoming over-confident during the later training process, likely to be the cause of actual forgetting*

Figure 4a and Figure 4b provides a visual representation of the entropy and accuracy trajectories over the instruction tuning process. The data manifests a consistent decrement in overall entropy, indicative of the model's escalating certainty in its predictions. Focusing on the late training process, we find that the entropy on both benchmark no longer fluctuate and undergoes a decline trend. Such results suggest that as the model's confidence in its answers intensifies, the precision of its responses to these questions concomitantly deteriorates. This phenomenon points towards overconfidence as a contributing factor to the emergence of actual forgetting.

**Finding3.** *LLMs are memorizing data throughout the instruction tuning phase, these data remains even after restoring the original pattern*

We can conclude from Figure 5 that as the instruction tuning process proceeds, LLMs start to memorize more instruction data and form a ascending trend along with the training process. Figure 3b shows that after the pattern is restored to its original pattern, a distinct gap is still observed from the base model, indicating the presence of actual forgetting. We can initially conclude from the ascending trend in Figure 5 and its comparison with 3b that when the data restored in LLMs exceed a certain threshold, the original knowledge distribution is affected, which cause the actual forgetting phenomenon.

## 6 Forgetting across models and data

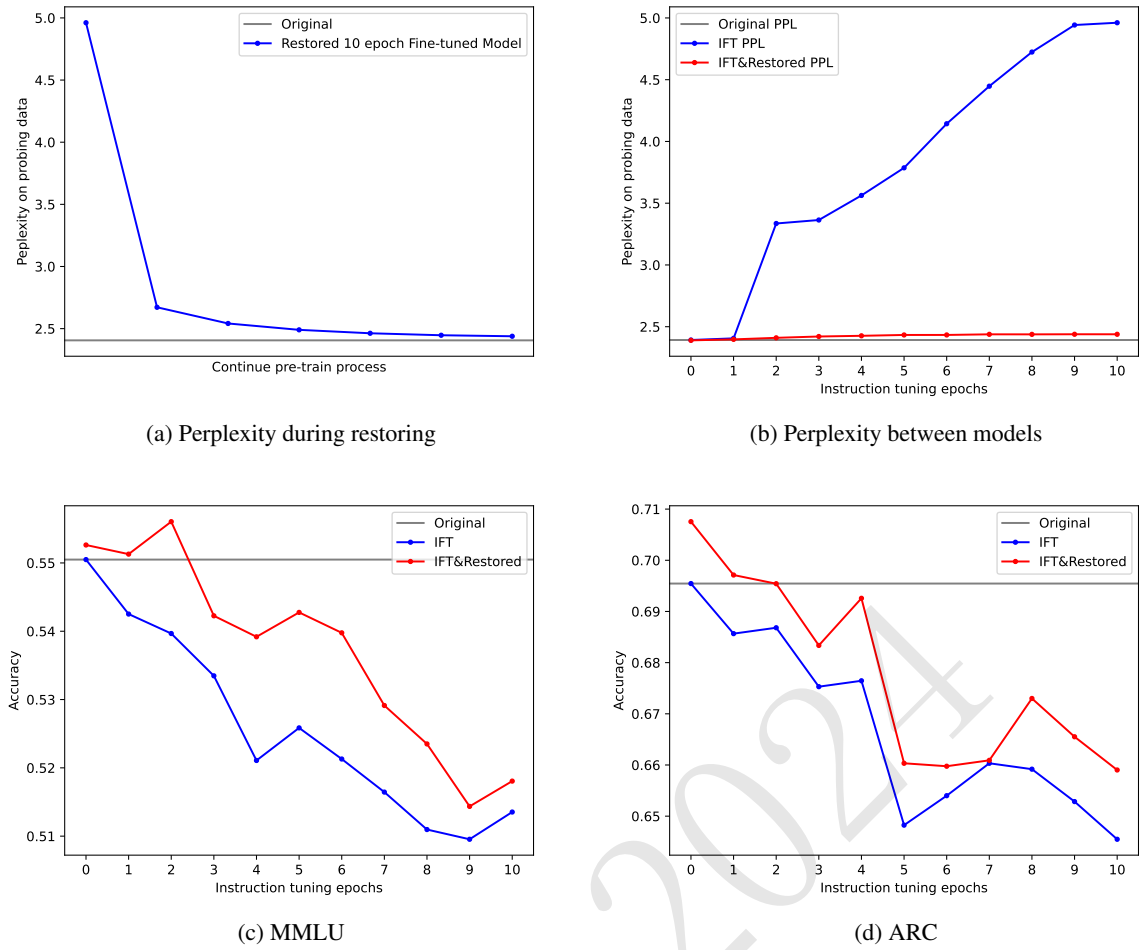
This section expands the forgetting phenomenon to a larger LLMs and other IFT data to investigate whether such a pattern of forgetting preserves across models. Experiment results implies a similar stage transition of forgetting, which helps to settle down the finding of this paper. The comparison with the smaller model reveals that larger model are more likely to face actual forgetting.

### 6.1 Result and Findings

**Finding 1.** *Pseudo forgetting and actual forgetting are also discovered across models*

Figure 6 illustrates both the perplexity and performance of LLaMa2-13b during the IFT process on alpaca as well as its restored version, from Figure 6a and Figure 6b we can conclude that the text pattern is restored which confirms the effectiveness of restoring stage on larger models. Figure 6c and Figure 6d

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1381-1394, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China




(a) Perplexity during restoring

(b) Perplexity between models

(c) MMLU

(d) ARC

Figure 6: Perplexity trajectory of LLaMa2-13b as well as the performance on knowledge benchmarks.

both indicate the existence of pseudo forgetting and actual forgetting with the gap between the original performance. We can conclude that the transition of forgetting keeps its consistency across the different LLMs, while the pattern shifting after IFT could also be restored to its original pattern.

**Finding 2.** *Larger Models are more likely to face actual forgetting*

Figure 6c and Figure 6d where we find that after two epochs of instruction fine-tuning, model fails to reach its original performance even if the text pattern is restored, While the actual forgetting in Figure 3c and Figure 3d begins after four epochs of instruction fine-tuning. Such difference indicate that larger models are more likely to lose their knowledge during instruction fine-tuning and face actual forgetting.

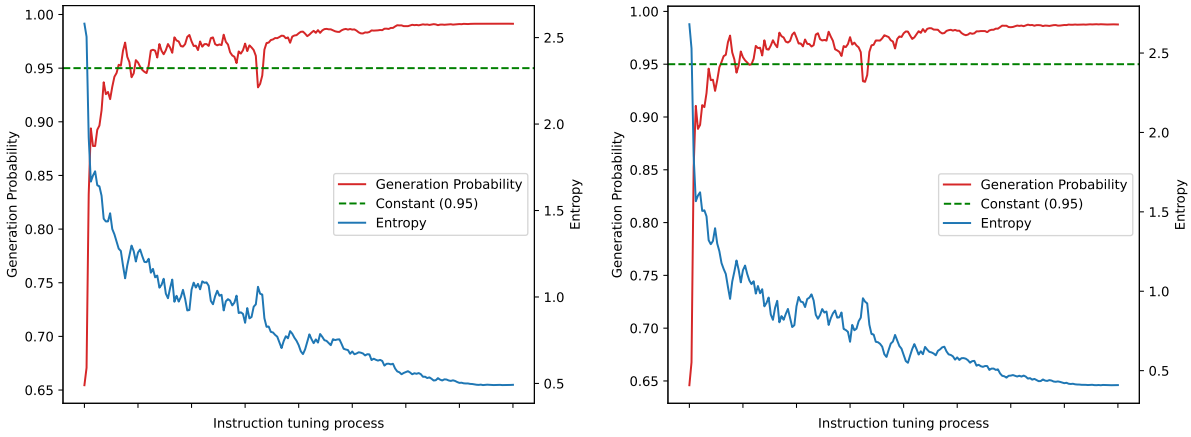
**Finding 3.** *format alignment and knowledge alignment persists across IFT data*

we rerun the testbed on LLaMa2-7b fine-tuned with vicuna and calculate the entropy and generation probability on output. From figure 7a and 7b we can also find that the generation probability of correct format answer also conveys at a high point, indicating the format alignment,while the fluctuation of entropy reveals the existence of knowledge alignment indicating that the specific characteristic also exist in models trained with other IFT data.

## 7 Conclusion

In our research, we have delved into the phenomenon of catastrophic forgetting within the context of Iterative Fine-Tuning (IFT) with the objective of elucidating its underlying mechanisms. Through our investigations, we have identified two distinct stages of forgetting: Pseudo Forgetting and Actual

CCL 2024




(a) MMLU (b) ARC

Figure 7: Entropy and generation probability across different selection in the vocabulary, where we find the format alignment finish during the pseudo forgetting, while the entropy descends as the training process goes on.

Forgetting. During the initial phase of IFT, we observe what we term as Pseudo Forgetting, where Large Language Models (LLMs) like LLaMa undergo a pattern shift rather than actual knowledge loss. The models adapt their existing patterns to new instructions while retaining the knowledge acquired during pre-training. This stage is characterized by a performance trajectory that initially oscillates before trending downward. In contrast, Actual Forgetting emerges in the later stages of IFT, signifying a genuine loss of knowledge within the LLMs. Our experiments demonstrate that models saved at later checkpoints exhibit a significant performance disparity when compared to their original pre-trained state, indicating that knowledge has indeed been forgotten during the instruction tuning process.

Our study also indicates the characteristic of these forgetting stages, LLMs are undergoing format alignment during the initial training process since the generation probability of correct format answer reaches a high point in just a few steps. After the completion of format alignment, the generation probability persist while the entropy across these correct format answer fluctuate, indicating the knowledge alignment. During the late stage of training, the entropy declines continuously, implying that models are over confident at their answer, which may cause the actual forgetting. Finally, by calculating the EM of LLMs on instruction data, we find that LLMs are memorizing IFT data throughout the instruction tuning process and may cause the actual forgetting.

In conclusion, our research provides significant insights into the stages of catastrophic forgetting in LLMs during IFT and establishes a foundation for future work aimed at mitigating such forgetting in Large Language Models.

## Limitations

Due to the constraint of computational resources, our major experiment was conducted on LLaMa, which may not cover all the forgetting mechanism throughout other model families, in the future we will extend our experiments to investigate the forgetting of instruction fine-tuning on other model families as well. Besides, owing to the lack of fair and intuitive metric, we apply multiple choice question as the probing methods for knowledge instead of generation question, we will continue to design a metric to investigate the forgetting phenomenon under the circumstances of generation.

## Acknowledgement

We sincerely thank the reviewers for their insightful comments and valuable suggestions. This research work is supported by the National Natural Science Foundation of China under Grants no. 62122077, 62106251 and 62306303.

## References


Jakob Smedegaard Andersen and Walid Maalej. 2022. Efficient, uncertainty-based moderation of neural networks text classifiers. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1536–1546, Dublin, Ireland. Association for Computational Linguistics.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeff Wu. 2023. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision.

Hyuntak Cha, Jaeho Lee, and Jinwoo Shin. 2021a. Co2l: Contrastive continual learning. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 9516–9525.

Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. 2021b. Swad: Domain generalization by seeking flat minima.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Gerhard Fischer. 2000. Lifelong learning—more than training. *Journal of Interactive Learning Research*, 11(3):265–294.

Robert M French. 1999. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135.

Sachin Goyal, Ananya Kumar, Sankalp Garg, Zico Kolter, and Aditi Raghunathan. 2022. Finetune like you pretrain: Improved finetuning of zero-shot vision models.

Prakhar Gupta, Cathy Jiao, Yi-Ting Yeh, Shikib Mehri, Maxine Eskenazi, and Jeffrey P. Bigham. 2022. Instructdial: Improving zero and few-shot generalization in dialogue through instruction tuning.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding.

Samyak Jain, Robert Kirk, Ekdeep Singh Lubana, Robert P. Dick, Hidenori Tanaka, Edward Grefenstette, Tim Rocktäschel, and David Scott Krueger. 2023. Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Suhas Kotha, Jacob Mitchell Springer, and Aditi Raghunathan. 2023. Understanding Catastrophic Forgetting in Language Models via Implicit Inference. ArXiv:2309.10105 [cs].

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. 2022. Fine-tuning can distort pretrained features and underperform out-of-distribution.

Po-Nien Kung and Nanyun Peng. 2023. Do Models Really Learn to Follow Instructions? An Empirical Study of Instruction Tuning. ArXiv:2305.11383 [cs].

Yong Lin, Lu Tan, Hangyu Lin, Zeming Zheng, Renjie Pi, Jipeng Zhang, Shizhe Diao, Haoxiang Wang, Han Zhao, Yuan Yao, and Tong Zhang. 2023. Speciality vs Generality: An Empirical Study on Catastrophic Forgetting in Fine-tuning Foundation Models. ArXiv:2309.06256 [cs].


CCL 2024

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2023. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*.

James L McClelland, Bruce L McNaughton, and Randall C O'Reilly. 1995. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102(3):419.

Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier.

Tan John Chong Min. 2023. An approach to solving the abstraction and reasoning corpus (arc) challenge.

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt

Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Binghui Peng and Andrej Risteski. 2022. Continual learning: a feature extraction formalization, an efficient algorithm, and fundamental obstructions. *Advances in Neural Information Processing Systems*, 35:28414–28427.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. 2017. icarl: Incremental classifier and representation learning.

Thomas Scialom, Tuhin Chakrabarty, and Smaranda Muresan. 2022. Fine-tuned language models are continual learners. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6107–6122, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Daniel L Silver, Qiang Yang, and Lianghao Li. 2013. Lifelong machine learning systems: Beyond learning algorithms. In *2013 AAAI spring symposium series*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Kushal Tirumala, Aram H. Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. 2022. Memorization Without Overfitting: Analyzing the Training Dynamics of Large Language Models. ArXiv:2205.10770 [cs].

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. LLaMA: Open and Efficient Foundation Language Models. ArXiv:2302.13971 [cs].

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open Foundation and Fine-Tuned Chat Models. ArXiv:2307.09288 [cs].

Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. 2023. Huatuo: Tuning llama model with chinese medical knowledge.

Han Wu, Haochen Tan, Kun Xu, Shuqi Liu, Lianwei Wu, and Linqi Song. 2022. Zero-shot cross-lingual conversational semantic role labeling. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 269–281, Seattle, United States. Association for Computational Linguistics.

CCL 2024

Mengzhou Xia, Mikel Artetxe, Chunting Zhou, Xi Victoria Lin, Ramakanth Pasunuru, Danqi Chen, Luke Zettlemoyer, and Ves Stoyanov. 2023. Training trajectories of language models across scales.

Lakshmi Narayana Yaddanapudi. 2016. The american statistical association statement on p-values explained. *Journal of anaesthesiology, clinical pharmacology*, 32(4):421.

Fan Yin, Jesse Vig, Philippe Laban, Shafiq Joty, Caiming Xiong, and Chien-Sheng Jason Wu. 2023. Did you read the instructions? rethinking the effectiveness of task definitions in instruction learning.

Shenglai Zeng, Yaxin Li, Jie Ren, Yiding Liu, Han Xu, Pengfei He, Yue Xing, Shuaiqiang Wang, Jiliang Tang, and Dawei Yin. 2023. Exploring memorization in fine-tuned language models.

Michael Zhang and Christopher Ré. 2022. Contrastive adapters for foundation model group robustness.

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, Alban Desmaison, Can Balioglu, Pritam Damania, Bernard Nguyen, Geeta Chauhan, Yuchen Hao, Ajit Mathews, and Shen Li. 2023. Pytorch fsdp: Experiences on scaling fully sharded data parallel.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

CCL 2024

# Prior Constraints-based Reward Model Training for Aligning Large Language Models

**Hang Zhou**[1], **Chenglong Wang**[1], **Yimin Hu**[1], **Tong Xiao**[1,2]*,
**Chunliang Zhang**[1], **and Jingbo Zhu**[1,2]
[1]NLP Lab, School of Computer Science and Engineering,
Northeastern University, Shenyang, China
[2]NiuTrans Research, Shenyang, China
{stceum, clwang1119}@gmail.com,
{xiaotong, zhujingbo}@mail.neu.edu.cn

## Abstract

Reinforcement learning with human feedback for aligning large language models (LLMs) trains a reward model typically using ranking loss with comparison pairs. However, the training procedure suffers from an inherent problem: the uncontrolled scaling of reward scores during reinforcement learning due to the lack of constraints while training the reward model. This paper proposes a **P**rior **C**onstraints-based **R**eward **M**odel (**PCRM**) training method to mitigate this problem. PCRM incorporates prior constraints—specifically, length ratio and cosine similarity between outputs of each comparison pair—during reward model training to regulate optimization magnitude and control score margins. We comprehensively evaluate PCRM by examining its rank correlation with human preferences and its effectiveness in aligning LLMs via RL. Experimental results demonstrate that PCRM significantly improves alignment performance by effectively constraining reward score scaling. As another bonus, our method is easily integrated into arbitrary rank-based alignment methods, such as direct preference optimization, and can yield consistent improvement. The code is available at https://github.com/wangclnlp/DeepSpeed-Chat-Extension/tree/PCRM.

## 1 Introduction

Reinforcement learning with human feedback (RLHF) has been proven to be an advanced technology to align large language models (LLMs) with human preferences (Ouyang et al., 2022; Ji et al., 2023; Wang et al., 2023b). It builds upon preference data, which rates and compares different outputs given the same input, where this rating is conducted by either human annotators or LLMs (Ouyang et al., 2022; Lee et al., 2023; Cui et al., 2023; Dubois et al., 2024). In practice, RLHF trains a reward model on the preference data with ranking loss for higher scores on preferred outputs than dispreferred ones. Then, RL algorithms such as Proximal Policy Optimization (PPO) (Schulman et al., 2017) are employed to fine-tune the LLM with the aim of optimizing this reward. During RL training, the reward model will give scores as signals, and the LLM will align the preference by increasing the probabilities of sampled outputs with higher reward scores. It is discernible that the alignment of the LLM is significantly influenced by how well the reward model is trained.

However, the training procedure of the reward model actually suffers from a failure mode: it learns from the preference data with the standard ranking loss yet provides a reward score for the sampled outputs during RL training. This training mode causes an inherent problem: the ranking loss increases the score margin between the outputs of this comparison without constraint during the training procedure, which makes an uncontrollable scale of scores in the process of RL training. For instance, the reward model trained by ranking loss can predict reward scores with a relatively large margin to two sampled outputs that do not differ very much (Zhu et al., 2023).

To address the problem, we propose a **P**rior **C**onstraints-based **R**eward **M**odel (**PCRM**) training method in this paper, which incorporates prior constraints while training the reward model. Specifically,

*Corresponding author.
©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1395-1407, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers

we select two features of the preference data for designing prior constraints: length ratio and cosine similarity of the outputs with the same prompt. The length ratio is computationally less demanding and directly reflects data differences, while the cosine similarity captures deeper semantic features despite being more computationally intensive. These constraints effectively regulate the optimization magnitude across different outputs for the same input, thereby controlling the margin of scores predicted by the reward model.

We comprehensively evaluate the proposed PCRM in the following two ways. Firstly, we test the rank correlation between the preferences predicted by PCRM and human preferences. It can assess the extent to which the reward model can serve as a surrogate for human-derived preference signals. Secondly, we verify the effectiveness of the PCRM in aligning an LLM via RL. This also demonstrates the influence of constraints on alignment performance straightly. Notably, PCRM can yield a +2.48% improvement in the GPT-4 win rate for the dialogue task compared to the traditional RLHF. Furthermore, we integrate our method into direct preference optimization (DPO), a rank-based alignment method. The results show that our method can also be effective in improving the rank-based alignment methods, *e.g.,* a 2.95% increase in the GPT-4 win rate on the dialogue task compared to DPO.

## 2 Related Work

Reinforcement learning with human feedback (RLHF) is a crucial technique to ensure that the behaviours of large language models (LLMs) are consistent with human preferences (Stiennon et al., 2020; Ouyang et al., 2022; Wang et al., 2023b). Recent works resorted to building a better reward model to improve the performance of RLHF on LLMs. These methods could be classified into three groups. The first group aimed to efficiently produce human preference data (Dubois et al., 2024; Cui et al., 2023; Lee et al., 2023). For instance, Lee et al. (2023) directly employed an LLM to annotate the comparison pairs. Significantly, the reward model trained on the LLM-annotated comparison pairs can achieve closer performance when applied to RLHF than the one trained on human-annotated comparison pairs. The second group tended to design fine-grained reward models to provide multiple reward scores for different reward criteria (Cheng et al., 2023; Wang et al., 2023b; Wu et al., 2024; Zhong et al., 2024). Notably, Wang et al. (2023b) learned various evaluation models from an LLM as reward models on the summarization task, including reward models scored for relevance, scored for fluency, scored for consistency, and scored for coherence. The third group that has attracted less attention generally explored how to merge multiple reward models in RLHF, such as reward model ensemble (Coste et al., 2023) and learning reward weights (Min et al., 2024). Different from these methods, in this paper, we proposed a superior training schema that enables this reward model to give a more accurate score in RLHF. Specifically, we refined the conventional ranking loss, typically utilized in reward model training, by incorporating prior constraints. These constraints are designed to regulate the margin between the scores of various outputs generated from the same input.

To the best of our knowledge, there is almost no previous work in training a reward model with prior constraints. The few related one is LLaMA2 (Touvron et al., 2023), which added a margin component in ranking loss for training each comparison pair. Although this margin simulates a constraint, it requires a manual annotation, which significantly increases the cost of training the reward model. Additionally, this work did not provide a specialized exploration or analysis on the topic of constraining the margin of scores during training reward models. In comparison, all of our prior constraints were computed automatically and did not require any human annotation. We also provided sufficient theoretical analysis and experiments to constrain the margin of scores. Furthermore, although there are some methods such as DPO (Amini et al., 2024) and RRHF (Yuan et al., 2023) that circumvent the need for training reward models, they still suffer from an inherent limitation from ranking loss, which is mentioned in DPO (Amini et al., 2024) without a solution. Our proposed PCRM can be easily extended to these methods to yield some benefits (See Section 5.7.2).

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1395-1407, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

## 3 Background

Human-preference alignment training is a key technique to ensure that the behaviours of LLMs are consistent with human preferences. Recent efforts to align LLMs have mainly been conducted via RLHF. It typically includes three stages: 1) collecting preference data, 2) training a reward model with ranking loss, and 3) optimizing an LLM against the reward model via RL.

### 3.1 Collecting Preference Data

The preference data consists of the given input $x$ and the corresponding different outputs sampled from LLMs trained by SFT (denoted as $\pi_\theta^{\text{SFT}}$) or annotated by humans, which we refer to as $(y_1, y_2, \cdots, y_n|x)$. In the preference data, the different outputs are rated and ranked by humans or LLMs with specific aspects (Ouyang et al., 2022; Dubois et al., 2024), which are denoted as $(y^{(1)} \succ y^{(2)} \succ \cdots \succ y^{(i)} \succ y^{(j)} \succ \cdots \succ y^{(n)}|x)$, where $y^{(1)}$ is the best while $y^{(n)}$ is the worst.

### 3.2 Training Reward Model

After preference data collection, we can train a reward model $\pi_\theta^{\text{RM}}$ from the preference data, where the training objective is to fit human preference. Note that we use $\pi_\theta^{\text{SFT}}$ to initialize the reward model. Suppose $r^*$ as the ideal model of human preference, based on the Bradley-Terry model (Bradley and Terry, 1952), the distribution of human preference $P_{\text{Human}}\left(y^{(i)} \succ y^{(j)}|x\right)$ can be written as:

$$P_{\text{Human}}\left(y^{(i)} \succ y^{(j)}|x\right) = \frac{\exp\left(r^*\left(y^{(i)}, x\right)\right)}{\exp\left(r^*(y^{(i)}, x)\right) + \exp\left(r^*(y^{(j)}, x)\right)} \tag{1}$$

$$= \sigma\left(r^*(y^{(i)}, x) - r^*(y^{(j)}, x)\right) \tag{2}$$

where $y^{(i)}$ denotes the data ranking $i$ and $\sigma$ denotes the Sigmoid activation function. When dealing with multiple outputs more than two, similarly, we can induce $P_{\text{Human}}(\cdot)$ based on the more general Plackett-Luce model (Plackett, 1975; Luce, 2005):

$$P_{\text{Human}}\left(y^{(1)} \succ y^{(2)} \succ \cdots \succ y^{(i)} \succ y^{(j)} \succ \cdots \succ y^{(n)}|x\right) = \prod_{i=1}^{n} \frac{\exp\left(r^*(y^{(i)}, x)\right)}{\sum_{j=i}^{n} \exp\left(r^*(y^{(j)}, x)\right)} \tag{3}$$

To learn the preference distribution, we need to increase the probability of preferred outputs. Here, it is typically achieved by a negative log-likelihood loss function:

$$\mathcal{L}_r = -\log P_{\text{Human}}\left(y^{(1)} \succ y^{(2)} \succ \cdots \succ y^{(i)} \succ y^{(j)} \succ \cdots \succ y^{(n)}|x\right) \tag{4}$$

$$= -\sum_{i=1}^{n} \log \frac{\exp\left(r^*\left(y^{(i)}, x\right)\right)}{\sum_{j=i}^{n} \exp\left(r^*\left(y^{(j)}, x\right)\right)} \tag{5}$$

Specially, when $n = 2$ which means there is only a comparison pair, the loss would be:

$$\mathcal{L}_r = -\log P_{\text{Human}}\left(y^{(i)} \succ y^{(j)}\right) \tag{6}$$

$$= -\log \sigma\left(r^*\left(y^{(i)}, x\right) - r^*\left(y^{(j)}, x\right)\right) \tag{7}$$

### 3.3 RL Training Against the Reward Model

In the process of RL training, we use the reward output $\pi_\theta^{\text{RM}}$ as signals, combined with an RL algorithm. Taking PPO as an instance, the corresponding loss for this training sample is given by:

$$\max_{\pi_\theta} \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta}\left[\pi_\theta^{\text{RM}}(x, y)\right] - \beta\mathbb{D}_{\text{KL}}\left[\pi_\theta||\pi_\theta^{\text{SFT}}\right] \tag{8}$$

where $\mathcal{D}$ is the dataset of RL training, $x$ is the input, and $y$ for the sampled outputs. $\mathbb{D}_{\text{KL}}$ is the KL dispersion which measures the distributional difference between $\pi_\theta^{\text{SFT}}$ and $\pi_\theta$, multiplied by $\beta$ which controls their distance, as the bigger $\beta$ is, the more significant constraint is applied.

CCL 2024

## 4 Prior Constraints-based Reward Model

Motivated by the challenges posed by the uncontrollable scale of scores in the reward model during the RL training process, our aim is to constrain the score margin between the outputs of this comparison when training the reward model. We propose the PCRM method to achieve this. Unlike conventional ranking loss, the proposed PCRM can constrain the maximum score margin between the outputs of each comparison with the length and cosine similarity features. In the following subsections, we will describe our PCRM in detail.

### 4.1 Optimization Objective of PCRM

Given the input $x$ and the reward model $\pi_\theta^{\text{RM}}$, the probability of $y_1 \succ y_2$ can be written as:

$$
\begin{aligned}
& P_{\pi_\theta^{\text{RM}}}(y_1 \succ y_2|x) && (9)\\
= \; & P_{\pi_\theta^{\text{RM}}}(y_1 - y_2 \succ 0|x) && (10)\\
= \; & \sigma\left(\pi_\theta^{\text{RM}}(y_1, x) - \pi_\theta^{\text{RM}}(y_2, x)\right) && (11)\\
= \; & \sigma\left(\Delta_{\pi_\theta^{\text{RM}}}(y_1, y_2, x)\right) && (12)
\end{aligned}
$$

where $\Delta_{\pi_\theta^{\text{RM}}}(\cdot)$ denotes the margin in the evaluation scores predicted by the reward model $\pi_\theta^{\text{RM}}$. If the reward model learns this probability with a standard ranking loss, the reward score of the preferred output will increase; conversely, the reward score of the dispreferred output will decrease. This endeavour aims to maximize the margin between the score of the preferred and dispreferred outputs as much as possible. However, in RLHF, we do not just expect the reward model to be able to distinguish which output is more preferred, but also to be able to give information on *how much* more one input is preferred. Consequently, we conjecture that facilitating the reward model to learn an appropriate score margin across different outputs could improve the performance of RLHF.

We achieve this goal by adding a maximum margin constraint, denoted as $\Delta^*(\cdot)$, where it takes values in the range $(0, +\infty)$. We re-derive Equation 9 with the constraint $\Delta^*(\cdot)$ as follows:

$$
\begin{aligned}
& P_{\pi_\theta^{\text{RM}}}(\Delta^*(y_1, y_2, x) \succ y_1 - y_2 \succ 0|x) && (13)\\
= \; & P_{\pi_\theta^{\text{RM}}}\left(\Delta^* > \Delta_{\pi_\theta^{\text{RM}}} > 0\right) && (14)\\
= \; & P_{\pi_\theta^{\text{RM}}}\left(\Delta^* > \Delta_{\pi_\theta^{\text{RM}}}\right) \times P_{\pi_\theta^{\text{RM}}}\left(\Delta_{\pi_\theta^{\text{RM}}} > 0\right) && (15)\\
= \; & \sigma\left(\Delta^* - \Delta_{\pi_\theta^{\text{RM}}}\right) \times \sigma\left(\Delta_{\pi_\theta^{\text{RM}}}\right) && (16)
\end{aligned}
$$

Based on the above derivation, we have the negative log-likelihood loss similar to the vanilla one:

$$
\begin{aligned}
\mathcal{L}_{\text{PCRM}} \; = \; & -\log P_{\pi_\theta^{\text{RM}}}\left(\Delta^* > \Delta_{\pi_\theta^{\text{RM}}} > 0\right) && (17)\\
= \; & -\log \sigma\left(\Delta^* - \Delta_{\pi_\theta^{\text{RM}}}\right) - \log \sigma\left(\Delta_{\pi_\theta^{\text{RM}}}\right) && (18)
\end{aligned}
$$

We define the maximum margin constraint through a negative correlation with the similarity between $y_1$ and $y_2$:

$$
\Delta^*(y_1, y_2, x) = \frac{\beta_1}{\text{Sim}(y_1, y_2, x) + \beta_2} + \beta_3 \tag{19}
$$

where $\text{Sim}(\cdot)$ denotes the similarity of the different outputs. $\beta_1$ controls the magnitude of the $\Delta^*(\cdot)$, $\beta_2$ controls the range of variation and $\beta_3$ controls the offset. We employ length ratio and cosine similarity to estimate $\text{Sim}(\cdot)$. When using the length ratio to estimate $\text{Sim}(\cdot)$, $\text{Sim}(\cdot)$ can given by:

$$
\text{Sim}_{\text{len\_rat}}(y_1, y_2, x) \; = \; \frac{\min(\phi(y_1), \phi(y_2))}{\max(\phi(y_1), \phi(y_2))} \tag{20}
$$

CCL 2024

where $\phi(y)$ denotes the length of $y$. Note that we do not need the input $x$ to be involved when employing the length ratio to estimate the $\mathrm{Sim}(\cdot)$. When considering the cosine similarity as another estimation of the $\mathrm{Sim}(\cdot)$, written as:

$$\mathrm{Sim}_{\text{cos\_sim}}(y_1, y_2, x) = 1 - \frac{1}{\pi} \times \arccos\left(\frac{E(x, y_1) \cdot E(x, y_2)}{\max(||E(x, y_1)||_2, \epsilon) \cdot \max(||E(x, y_2)||_2, \epsilon)}\right) \quad (21)$$

where $\epsilon$ is a small value to avoid division by zero and $\arccos(\cdot)$ is the inverse of the cosine function. Inspired by the strong text encoding capability of the pre-trained model (Devlin et al., 2018; Xiao and Zhu, 2023), we employ a pre-trained model like BERT to compute $E(\cdot)$. There are other choices to define $\mathrm{Sim}(\cdot)$ for specific tasks. For instance, we can use the ROUGE function (Lin, 2004) to define $\mathrm{Sim}(\cdot)$ in the summarization task.

### 4.2 Analysis of the Optimization for PCRM

To better understand the constrained optimization for PCRM, we take the derivative of the loss function. The gradient of $\mathcal{L}_{\text{PCRM}}$ with respect of the parameters $\theta$ is:

$$\nabla_\theta \mathcal{L}_{\text{PCRM}} = \nabla_\theta \Delta_{\pi_\theta^{\text{RM}}} \times \left(\sigma\left(\Delta_{\pi_\theta^{\text{RM}}} - \Delta^*\right) - \sigma\left(-\Delta_{\pi_\theta^{\text{RM}}}\right)\right) \quad (22)$$

As a comparison, we also take the derivative of one for vanilla reward loss, which can be written as:

$$\nabla_\theta \mathcal{L}_{\text{RM}} = \nabla_\theta \Delta_{\pi_\theta^{\text{RM}}} \times \left(-\sigma\left(-\Delta_{\pi_\theta^{\text{RM}}}\right)\right) \quad (23)$$

Compare Equation 22 with 23, we can find their difference $\sigma\left(\Delta_{\pi_\theta^{\text{RM}}} - \Delta^*\right)$ which is always positive and will decrease the coefficient of $\nabla_\theta \Delta_{\pi_\theta^{\text{RM}}}$, constraining the optimization thereby.

When $\sigma\left(\Delta_{\pi_\theta^{\text{RM}}} - \Delta^*\right) - \sigma\left(-\Delta_{\pi_\theta^{\text{RM}}}\right) = 0$ in Equation 22, it implies that $\Delta_{\pi_\theta^{\text{RM}}} - \Delta^* = -\Delta_{\pi_\theta^{\text{RM}}}$ due to the monotonic increase of the Sigmoid activation function. Thus, we can deduce $\Delta_{\pi_\theta^{\text{RM}}} = \frac{\Delta^*}{2}$. Then we can get the conclusion that when $\Delta_{\pi_\theta^{\text{RM}}} < \frac{\Delta^*}{2}$, the coefficient of $\nabla_\theta \Delta_{\pi_\theta^{\text{RM}}}$ has the same sign with the one of the vanilla reward gradient, meaning the same optimization direction with the origin; in contrast, when $\Delta_{\pi_\theta^{\text{RM}}} > \frac{\Delta^*}{2}$, there will be opposite optimization direction decreasing the margin of reward scores. In this way, we can control the distance of scores of different outputs while optimizing the reward model.

## 5 Experiments

We evaluate the proposed PCRM in the following two ways. Firstly, we test the reward model trained by PCRM. Secondly, we analyze the performance of applying this trained reward model to RLHF. We conduct experiments on the commonly used generation tasks, including dialogue and summarization.

### 5.1 Datasets

The datasets used for each task are as follows:

- *Dialogue*: We employed AlpacaFarm (Dubois et al., 2024) dataset on dialogue task, which consists of 10K supervised fine-tuning split, 10K pairwise preference split, 20K unlabeled split for RL training, and 2K validation split based on 52k Alpaca data (Taori et al., 2023). For evaluation, we employed their evaluation set, which contains 805 instructions selected from a series of open-source datasets with real-world user interactions as reference instructions.

- *Summarization*: We used the filtered versions of the TL;DR dataset and human feedback dataset provided by OpenAI (Stiennon et al., 2020) on summarization task, the former one for instruction tuning and alignment, and the latter one for reward modeling. The TL;DR dataset is filtered to ensure quality and contains 123.2K samples in the final version, including 116.7K for training, 6.4K for validating, and 6.5K for testing. The large, high-quality preference dataset of human comparisons between summaries contains a 92.9K training set, a 33.1K validation set, and a 50.7K test set. Due to the enormous computational cost caused by the vast testing set, we randomly selected 10% as our final test set.

CCL 2024

### 5.2 Settings

| Task | Sim(·) | $\beta_1$ | $\beta_2$ | $\beta_3$ | max_len |
|---|---|---|---|---|---|
| Dialogue | len_rat | 10 | 0.001 | -5 | 512 |
| | cos_sim | 20 | 0.001 | -15 | |
| Summarization | len_rat | 10 | 0.001 | -5 | 1024 |
| | cos_sim | 20 | 0.001 | -15 | |

Table 1: The hyper-parameters used in our experiments.

We conducted our experiments based on DeepSpeed-Chat[1] with a cross-entropy loss on supervised fine-tuning, where prompt parts were masked, and a ranking loss was employed by reward modeling. We set the maximum sequence length to 512 and 1024 for the dialogue and summarization tasks, respectively. For calculating the similarity between pair-wise data for setting prior constraints, we employed the BERT-base model (Devlin et al., 2018) to encode paired data and then calculated the similarity between `[CLS]` embeddings. Considering that the maximum sequence length supported by BERT is 512; however, the sequence length is 1024 for summarization, we set the tokenizer to truncate from the left, *i.e.,* truncating part of the prompt and calculating the semantic similarity of the remaining part. Additionally, the value of $\beta_1$, $\beta_2$ and $\beta_3$ mentioned in Eq. 19 are shown in Table 1. The performance of other hyper-parameters is reported in Section 5.7.1. Furthermore, we employed the top-$p$ sampling method in the process of generation, where the temperature and the $p$ were set to 0.75 and 0.95, respectively.

### 5.3 Evaluation Metrics

We evaluated the effectiveness of our method comprehensively from two dimensions: training reward models and aligning LLMs with the trained reward models. For training reward models, we scored the pair-wise test set. Based on this, we calculated the accuracy of the predicted scores (for details of calculating accuracy, see Appendix A). We trained three epochs on the training set and saved the model for each epoch. We selected the best one with the validation set. We compared the performance of the vanilla method with ours, as shown in the following section. To align LLMs with the trained reward models, we used different metrics depending on the tasks. Specifically, for the dialogue task, we used PandaLM (Wang et al., 2023c) and GPT-4 to choose a better one from the model output and the reference and calculate the win rate following Rafailov et al. (2024). For the summarization task, except for GPT-4, we also employ ROUGE (Lin, 2004) and BARTScore (Yuan et al., 2021) to evaluate the quality of the summary generated by the model from the posts.

### 5.4 Results of Training Reward Models

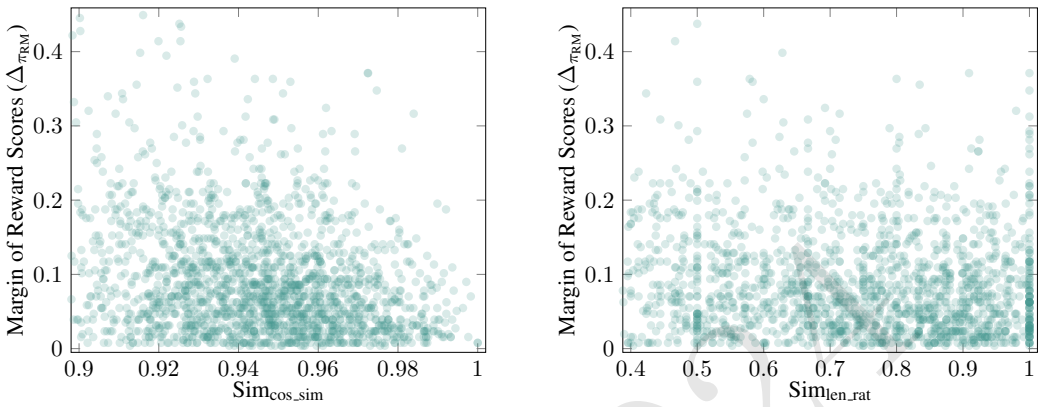
The performance of the PCRM on the dialogue and the summarization task are presented in Table 2. We can see that the accuracy of the PCRM increases slightly, either by calculating similarity with cosine similarity or length ratio. As a comparison, we test the reward models trained with random and fixed constraints, whose performance is close to the vanilla one with some decrease. This result demonstrates the importance of the prior information. The appropriate value of the constraints also affects the performance because the training object of PCRM can be divided into two parts: achieving higher accuracy and controlling the distribution of the reward scores. The latter part may conflict with the former under extreme constraints (See Section 5.7.1).

To further explore the distribution of reward scores, we visualize the relationship between the margin of predicted reward scores and the similarity of the paired data on dialogue task, with or without constraints in Figure 1. The area with deeper colour in the figure represents more points gathering at that position. The points in Figure 1 (a) are more dispersed than the ones shown in Figure 1 (b) below, as the upper right part of the figures with constraints are more clean. From Figure 1 (b), we can observe that

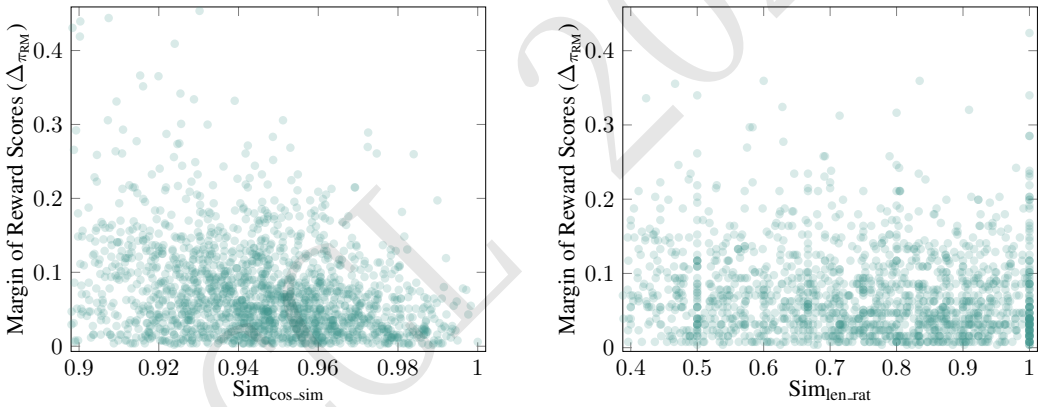
[1] https://github.com/microsoft/DeepSpeedExamples

| Method | Dialogue | Summarization |
|---|---|---|
| Vanilla RM | 54.93 | 72.20 |
| PCRM-Random | 54.13 | 71.70 |
| PCRM-Fixed | 54.40 | 71.70 |
| PCRM-Length (Ours) | 55.87 | 72.50 |
| PCRM-Cosine (Ours) | **56.53** | **72.60** |

Table 2: The accuracy of reward models on the dialogue and summarization task. "-Random" means using random constraints, "-Fixed" means using fixed one, "-Cosine" means using constraints calculated by cosine similarity, and "-Length" for length ratio.




(a) w/o constraints




(b) w/ constraints

Figure 1: The distribution between the margin of predicted reward scores and the similarity of paired data (calculated by cosine similarity or length ratio) with or without constraint on the dialogue task. Each green point corresponds to a single data sample. The x-axis refers to the similarity calculated by the cosine similarity of sentence embedding or by the ratio of sentence length. The y-axis refers to the margin of the reward scores.

the margin of those points is limited to a lower level because of their higher similarity. This observation confirms that our method can provide an effective constraint during the training of the reward model. We can draw similar observations on the summarization task (See Figure 3 in Appendix B).

### 5.5 Results of Dialogue

The results of our alignment with PCRM on downstream tasks, as detailed in Table 3, show that PCRM significantly outperforms the vanilla reward model. Specifically, when utilizing cosine similarity, PCRM surpasses RLHF by +3.18 points on PandaLM and +2.48 points on the GPT-4 win rate in dialogue task.

| Method | Dialogue | | Summarization | | |
|---|---|---|---|---|---|
| | PandaLM | GPT-4 Win | ROUGE-L | BARTScore | GPT-4 Win |
| SFT | 65.75 | 55.77 | 22.93 | -6.31 | 41.83 |
| RLHF | 72.73 | 60.13 | 25.10 | -5.17 | 70.53 |
| PCRM-Random | 70.81 | 59.75 | 24.03 | -6.02 | 50.53 |
| PCRM-Fixed | 69.34 | 60.71 | 24.87 | -6.09 | 52.21 |
| PCRM-Length (Ours) | 73.52 | 62.30 | 22.76 | **-4.90** | 71.87 |
| PCRM-Cosine (Ours) | **75.91** | **62.61** | **25.97** | -5.34 | **72.52** |

Table 3: The performance of alignment with PCRM on downstream tasks.

| $\beta_1$ | $\beta_3$ | Accuracy with $\text{Sim}_{\text{cos\_sim}}(\cdot)$ | Accuracy with $\text{Sim}_{\text{len\_rat}}(\cdot)$ |
|---|---|---|---|
| 10 | -3 | 54.79 | 55.87 |
| 10 | -5 | 55.73 | **55.87** |
| 10 | -7 | 53.92 | 55.33 |
| 10 | -9 | 52.26 | 51.60 |
| 20 | -13 | 53.73 | 55.33 |
| 20 | -15 | **56.53** | 55.73 |
| 20 | -17 | 54.80 | 55.47 |
| 20 | -19 | 55.20 | 55.20 |
| 30 | -25 | 54.80 | 55.60 |
| 40 | -35 | 55.60 | 55.07 |

Table 4: The performance of PCRM with different constraints measured by accuracy. The range and the numerical size of the constraints are controlled with different pairs of $\beta_1$, $\beta_3$, and $\beta_2$ is fixed to 0.001 for computational stability.

The importance of meaningful constraints cannot be overstated. Our tests with PCRM using both random and fixed constraints revealed that these constraints were either marginally effective or even detrimental to alignment progress. Furthermore, we found that prior constraints enriched with additional information yielded better results. This is evident from comparing the effectiveness of length ratio and cosine similarity in calculating similarity; the superior performance of cosine similarity can be attributed to its ability to capture implicit prior semantic information of sentences, whereas the length ratio approach is more superficial and offers less valuable information.

### 5.6 Results of Summarization

Except for the dialogue task, we achieve similar results with our PCRM method on the summarization task as shown in Table 3. Specifically, PCRM with cosine similarity outperforms RLHF by +0.87 points on ROUGE-L and +1.99 points on GPT-4 win rate. As to PCRM with length ratio, it outperforms RLHF by +0.27 points on BART Score and +1.34 points on GPT-4 win rate. It is obvious that there is a misalignment between ROUGE-L and BART Score- the models with high BART Score may not necessarily achieve high ROUGE-L scores. We attribute this phenomenon to the probable low correlation with human judgments, which is also reported in Wang et al. (2023a).

### 5.7 Analysis

#### 5.7.1 Performance of the Variety Range and Numerical Size of Constraints on PCRM

The variety range and numerical size of constraints are two key factors that control the maximum score margin and the strength of constraints. Therefore, we conduct experiments to study the impact of the

| Method | PandaLM | GPT-4 Win |
|---|---|---|
| SFT | 65.75 | 55.77 |
| DPO | 75.64 | 60.96 |
| PCDPO-Length (Ours) | 77.02 | 61.11 |
| PCDPO-Cosine (Ours) | **78.11** | **63.91** |

Table 5: The experiment results of PCDPO.

variety range and the numerical size of constraints. Specifically, we explore with $\beta_2$ fixed to 0.001 in Equation 19 for computational stability, and different $\beta_1$, $\beta_3$, for variety range and numerical size on PCRM. The results are summarized in Table 4. From the results, we can observe that the unsuitable constraints may hurt the performance. We conjecture that the larger one may weaken the effect of distribution control, and the smaller one may conflict with the typical optimization. Based on these experimental results, we find that the optimal $\beta_1$ and $\beta_3$ are 20 and -15 for the cosine similarity constraint; the optimal $\beta_1$ and $\beta_3$ are 10 and -5 for the length ratio constraint. Note that $\beta_2$ prevents the occurrence of a similarity $\mathrm{Sim}(\cdot)$ of 0, so we do not tune it and simply set a relatively small value.

### 5.7.2 Integrating PCRM to DPO

Rafailov et al. (2024) proposes a direct preference optimization (DPO) method that bypasses the reward modeling step and directly optimizes an LLM using preference data. However, this method still performs optimization based on ranking loss, which has the same limitation during training with preference data. Therefore, we attempt to integrate the proposed PCRM into DPO, constraining the direct preference optimization. Suppose we have the LLM trained by SFT $\pi_\theta^{\text{SFT}}$, and based on it, we fine-tune the model with DPO and preference data $(y_1, y_2, x) \sim \mathcal{D}$ where $y_1 \succ y_2$. Rafailov et al. (2024) points out that the reward model can be represented as follows with some mathematical transformation on the optimization objective of PPO:

$$\pi_\theta^{\text{RM}}(y, x) = \beta \log \frac{\pi_\theta^{\text{RM}}(y|x)}{\pi_\theta^{\text{SFT}}(y|x)} + \beta \log Z(x) \tag{24}$$

where $Z(x) = \Sigma_y \pi_\theta^{\text{SFT}}(y|x) \exp\left(\frac{1}{\beta}\pi_\theta^{\text{RM}}(y|x)\right)$. Bring the above equation into Equation 16 and 18, we can obtain the loss of **P**rior **C**onstraints-based **DPO** (**PCDPO**):

$$\begin{aligned}
\mathcal{L}_{\text{PCDPO}} &= -\log P_{\pi_\theta^{\text{RM}}}\left(\Delta^* > \Delta_{\pi_\theta^{\text{RM}}} > 0\right) && (25)\\
&= -\log \sigma\left(\Delta^* - \Delta_{\pi_\theta^{\text{RM}}}\right) - \log \sigma\left(\Delta_{\pi_\theta^{\text{RM}}}\right) && (26)\\
&= -\log \sigma\left(\Delta^* - \beta \log \frac{\pi_\theta^{\text{RM}}(y_1|x)}{\pi_\theta^{\text{SFT}}(y_1|x)} + \beta \log \frac{\pi_\theta^{\text{RM}}(y_2|x)}{\pi_\theta^{\text{SFT}}(y_2|x)}\right)\\
&\quad -\log \sigma\left(\beta \log \frac{\pi_\theta^{\text{RM}}(y_1|x)}{\pi_\theta^{\text{SFT}}(y_1|x)} - \beta \log \frac{\pi_\theta^{\text{RM}}(y_2|x)}{\pi_\theta^{\text{SFT}}(y_2|x)}\right) && (27)
\end{aligned}$$

With the theory above, we conduct experiments with the constraints on the alpaca dataset. The experimental results are shown in Table 5. From the results, we can find that PCRM can yield consistency improvements in DPO. Notably, when armed with the cosine similarity constraint, we can obtain a +2.95 points improvement on the GPT-4 win rate.

### 5.7.3 Inference with Different Temperature

It is known to us that temperature influences the performance of inference, so we make inferences with different sampling temperatures on the dialogue task three times and calculate their mean in case of random variations while comparing different methods. Results are reported in Figure 2.

The experiment results are consistent with the previous. Our method, PCRM with cosine similarity, outperforms the vanilla reward model no matter which sampling temperature is used, and the increase

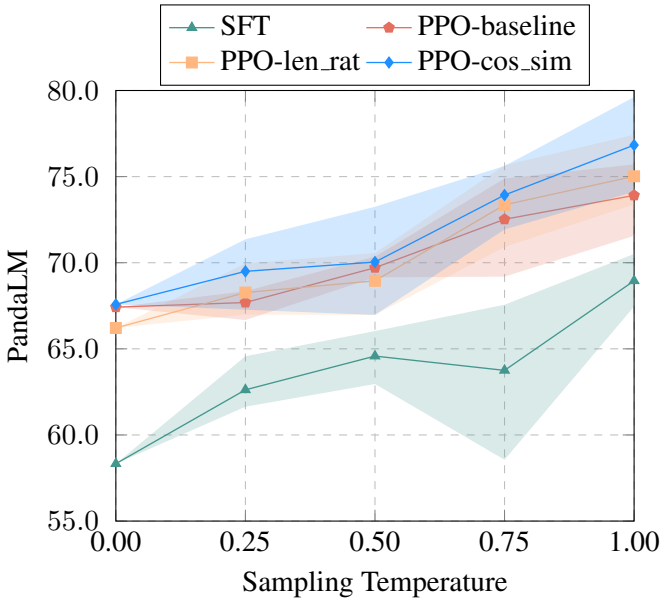



Figure 2: PandaLM scores for different sampling temperatures using different methods. For each dialogue model, we conduct the generation three times and report the mean score of these generated responses.

of PCRM with length ratio outperforms most of the temperatures. As analyzed before, we infer that the performance of the length ratio is limited due to the easy-to-learn prior information.

## 6 Conclusions

RLHF plays an important part in applications based on LLMs for their ability to align language models with human preferences. As a proxy for human preferences, the reward model makes a significant contribution. We have introduced PCRM, a method of training reward models with prior constraints to better control the distribution of reward scores without sacrificing prediction accuracy. Instead of optimizing the vanilla reward model with the goal of maximizing the margin of scores, PCRM restricts the optimization process with the provided prior information and fits the distribution of reward scores for better alignment. Furthermore, this method is applicable to DPO, which treats the language model as a reward model and optimizes it directly. Through experiments on downstream tasks, we have validated the effectiveness of this method.

## 7 Limitations

Our results raise several questions that are beyond the scope of the present study: How can we determine the appropriate range of constraints for different tasks or datasets? For example, we explore hyperparameters in dialogue and summarization tasks, what about when it comes to a brand new task? Can the function of the constraints be replaced with any other prior information, and if so, will that be effective? Finally, it would be advantageous if the prior constraints could be learned automatically from data, rather than being manually set.

## Acknowledgements

This work was supported in part by the National Science Foundation of China (No.62276056), the Natural Science Foundation of Liaoning Province of China (2022-KF-16-01), the Fundamental Research Funds for the Central Universities (Nos. N2216016 and N2316002), the Yunnan Fundamental Research Projects (No. 202401BC070021), and the Program of Introducing Talents of Discipline to Universities, Plan 111 (No.B16009).

## References


Afra Amini, Tim Vieira, and Ryan Cotterell. 2024. Direct preference optimization with an offset. *arXiv preprint arXiv:2402.10571*.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Pengyu Cheng, Jiawen Xie, Ke Bai, Yong Dai, and Nan Du. 2023. Everyone deserves a reward: Learning customized human preferences. *arXiv preprint arXiv:2309.03126*.

Thomas Coste, Usman Anwar, Robert Kirk, and David Krueger. 2023. Reward model ensembles help mitigate overoptimization. *arXiv preprint arXiv:2310.02743*.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2023. Ultrafeedback: Boosting language models with high-quality feedback. *arXiv preprint arXiv:2310.01377*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. 2024. Alpacafarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36.

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, et al. 2023. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

R Duncan Luce. 2005. *Individual choice behavior: A theoretical analysis*. Courier Corporation.

Do June Min, Veronica Perez-Rosas, Kenneth Resnicow, and Rada Mihalcea. 2024. Dynamic reward adjustment in multi-reward reinforcement learning for counselor reflection generation. *arXiv preprint arXiv:2403.13578*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Robin L Plackett. 1975. The analysis of permutations. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 24(2):193–202.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Chenglong Wang, Hang Zhou, Kaiyan Chang, Tongran Liu, Chunliang Zhang, Quan Du, Tong Xiao, and Jingbo Zhu. 2023a. Learning evaluation models from large language models for sequence generation. *arXiv preprint arXiv:2308.04386*.


CCL 2024

Chenglong Wang, Hang Zhou, Yimin Hu, Yifu Huo, Bei Li, Tongran Liu, Tong Xiao, and Jingbo Zhu. 2023b. Esrl: Efficient sampling-based reinforcement learning for sequence generation. *arXiv preprint arXiv:2308.02223*.

Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, et al. 2023c. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. *arXiv preprint arXiv:2306.05087*.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2024. Fine-grained human feedback gives better rewards for language model training. *Advances in Neural Information Processing Systems*, 36.

Tong Xiao and Jingbo Zhu. 2023. Introduction to transformers: an nlp perspective. *arXiv preprint arXiv:2311.17633*.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. 2023. Rrhf: Rank responses to align language models with human feedback without tears. *arXiv preprint arXiv:2304.05302*.

Yifan Zhong, Chengdong Ma, Xiaoyuan Zhang, Ziran Yang, Qingfu Zhang, Siyuan Qi, and Yaodong Yang. 2024. Panacea: Pareto alignment via preference adaptation for llms. *arXiv preprint arXiv:2402.02030*.

Banghua Zhu, Jiantao Jiao, and Michael I Jordan. 2023. Principled reinforcement learning with human feedback from pairwise or $k$-wise comparisons. *arXiv preprint arXiv:2301.11270*.

CCL 2024

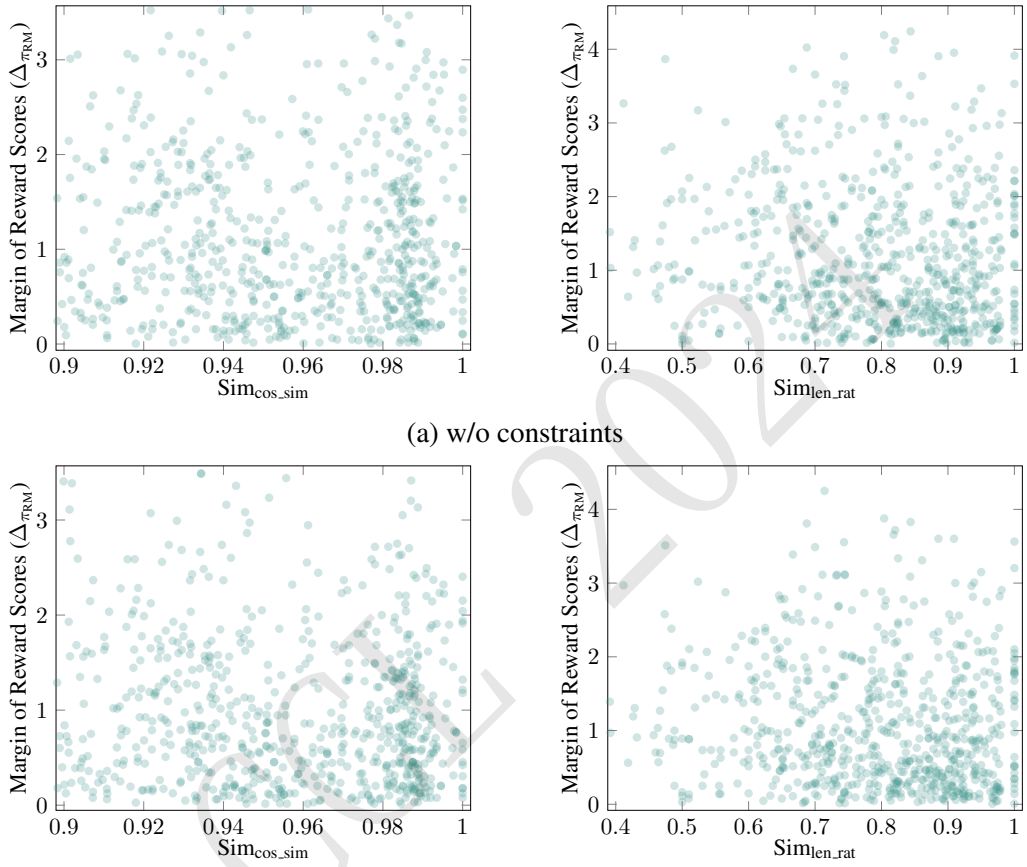
## Appendix A. Calculating Accuracy of the Reward Scores

Suppose we have a test set for human preference $(x, y_1, y_2) \sim \mathcal{D}_{\text{test}}$, in which $y_1$ is preferred than $y_2$ with the same $x$ by human, and the corresponding scores predicted by the reward model are $\pi_\theta^{\text{RM}}(y1, x)$, $\pi_\theta^{\text{RM}}(y2, x)$. The accuracy of the scores is defined as:

$$\text{Acc}(\pi_\theta^{\text{RM}}, \mathcal{D}_{\text{test}}) = \frac{\text{Count}_{(x,y_1,y_2)\sim\mathcal{D}_{\text{test}}}\left(\pi_\theta^{\text{RM}}(y1, x) > \pi_\theta^{\text{RM}}(y2, x)\right)}{\text{Count}\left((x, y_1, y_2) \in \mathcal{D}_{\text{test}}\right)} \tag{28}$$

where $\text{Count}(\cdot)$ denotes the total number of the samples that meet the condition.

## Appendix B. Distribution of Reward Scores for Summarization Task




(a) w/o constraints

(b) w/ constraints

Figure 3: The distribution between the margin of predicted reward scores and the similarity of paired data (calculated by cosine similarity or length ratio) with or without constraint on the summarization task.

CCL 2024

# Prompt Engineering 101
# Prompt Engineering Guidelines from a Linguistic Perspective

**Wenjuan Han**[1]*, **Xiang Wei**[1]*, **Xingyu Cui**[1]*, **Ning Cheng**[1]*,
**Guangyuan Jiang**[2], **Weinan Qian**[2] **Chi Zhang**
[1] Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing, China
[2] Peking University, Beijing, China
wjhan@bjtu.edu.cn, 22120436@bjtu.edu.cn
22120357@bjtu.edu.cn, ningcheng@bjtu.edu.cn
jgy@stu.pku.edu.cn, ypqwn@stu.pku.edu.cn, chi.zhang@ucla.edu

## Abstract

Deploying tuning-free prompting is challenging in engineering practice: it not only requires users to engage in cumbersome trials and errors but is also extremely time-consuming, as even a slight change in wording and phrasing could have a huge impact on the final performance. To further investigate the impact of different prompts, in this work, we perform a systematic inspection of four factors in linguistics involved in prompt engineering: syntax, semantics, lexicon, and pragmatics. The empirical results quantify the sensitivity of the output to small textual perturbations in four linguistic factors of prompts. Based on the analysis of these four factors, we present a series of design guidelines to help human users write effective prompts. Human evaluation on amateurs shows that using the proposed guidelines helps humans produce prompts with significant gains in zero-shot performance in Pre-trained Language Models (PLMs) and hence validates the utility of the guidelines.

## 1 Introduction

Dramatic gains in Large Language Models (LLMs) have been witnessed in recent years (Brown et al., 2020; Rae et al., 2021; Thoppilan et al., 2022; Chowdhery et al., 2022), accelerated by the discovery of prompting (Liu et al., 2021; Wei et al., 2022; Zhang et al., 2021). Prompts function in the form of natural language to drive LLMs to access a variety (Wei et al., 2023; Wei et al., 2024) of downstream tasks (See Sec. A). However, amateurs must engage in cumbersome trials and errors when the input prompt quality is poor or ambiguous for LLMs to deliver their real intention (Jiang et al., 2020; Lu et al., 2022; Kim et al., 2022; Madaan and Yazdanbakhsh, 2022; Khashabi et al., 2022). Guiding amateurs to clearly present their intention, we refer to education and linguistic theory where linguistics knowledge often be utilized to guide second-language students to effectively express ideas in writing. There is a vast body of research available in the field of linguistics that delves into these topics (Givón, 1978; Karakoç and Köse, 2017).

Inspired by this, we investigate ***linguistics*** factors of prompts. A general overview of how these linguistic factors can be employed to support humans is as follows. **Syntax** studies how words and morphemes combine to form larger phrases and gradually constitute a sentence (Lyons, 1981; Silva and Matsuda, 2012). Central concerns of syntax include word order, verb tense, subject-verb agreement, hierarchical sentence structure, and other syntactic features. Understanding syntax can aid second language students in composing coherent sentences (Ferris, 1999). **Semantics** studies the meaning in natural languages. Negation, denoting the reversal of some meaning groups in a sentence, is one of the most discussed phenomena in semantics (Givón, 1978). Human beings can learn about negation and can benefit from developing their understanding of the appropriate use of words and expressions in specific contexts. **Lexicon** or lexicology, studies the vocabulary of a language. (Johnson et al., 2016) and (Karakoç and Köse, 2017) examine

* Equal Contribution

©2024 China National Conference on Computational Linguistics
Published under Creative Commons Attribution 4.0 International License

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1408-1426, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

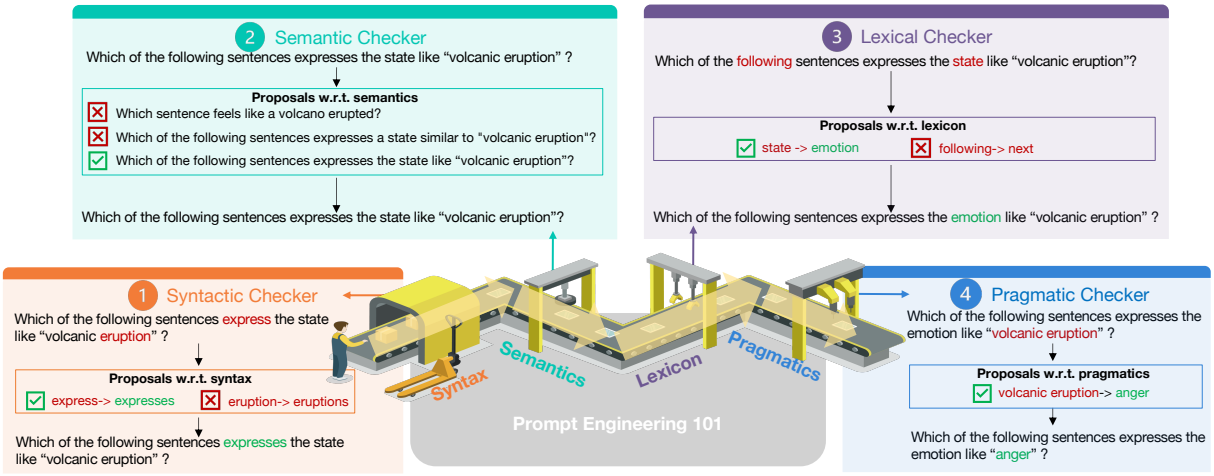



Figure 1: Illustration of our guidelines `Prompt Engineering 101`. The guidelines involve four stages: a syntactic checker to correct grammatical errors, a semantic checker to make proposals without negation, a lexical checker to substitute works, and a pragmatic checker to remove metaphors. `Prompt Engineering 101` is used for an amateur to interact with PLMs via prompting. Following this workflow, the amateur checks carefully each factor and finally writes a prompt. For better checking, in each stage, some suggestions (bullet points in the box) from external automatic software (*e.g.*, grammatical error corrector) will be provided for your reference. The amateur can choose to accept the suggestion or not.

the vocabulary knowledge and suggest effective vocabulary strategies for improving writing proficiency. Similarly, (Razeghi et al., 2022) consistently shows that LLMs are more accurate on samples with words more commonly used. **Pragmatics** is the study of using languages to express what we call "Beyond Saying" in linguistics and cognitive science (Korta and Perry, 2020). Generally speaking, semantics focuses on the literal meanings of linguistic expressions, whereas pragmatics captures the context-dependent implicit factors of meaning (Fried et al., 2022). The study of pragmatics involves metaphors, deceits, indirect speech, irony, Gricean maxims, humor, and coherence inferences (Hu et al., 2022). (Kasper and Roever, 2005) offers insights into how pragmatic knowledge helps produce appropriate sentences. Against this background, we explore the aforementioned four crucial factors: syntax, semantics, lexicon, and pragmatics, since these four factors provide valuable insights that can help humans produce coherent sentences and present desired intentions.

We systematically and quantitatively inspect the impact of the four linguistics factors on prompting LLMs (Sec. 2). For each factor, we study the impact on zero-shot performance via controlled experiments. For syntax, we randomly permute words and introduce grammatical errors in prompts in the experimental group (See Sec. 2.1). For semantics, we explore whether prompts with perturbed semantics but correct syntax can still produce meaningful outputs (See Sec. 2.2). For lexicon, we study prompts with different wording and phrasing. In particular, we replace words with synonyms or translate them into other languages, *e.g.*, Chinese (See Sec. 2.3). Lastly, we investigate the pragmatics of metaphors. We build QA tasks with and without metaphors and study their impact on performance (See Sec. 2.4). The results from these experiments show clear paths for us to navigate in the space of prompt design and steer LLMs toward desirable generations. We, therefore, present a series of design guidelines `Prompt Engineering 101` (Sec. 3) to help humans and the automated toolkit refine prompts. Evaluation on both humans (Sec. 4.2) and automated toolkit (Sec. 4.3) show that using the proposed guidelines subjects produce prompts with significant gains in zero-shot performance in LLMs and hence validates the utility of the guidelines.

CCL 2024

## 2 Analysis of impact of Linguistic Components

This section describes our analysis of the four linguistic components and the proposed guidelines to operationalize them. We list the guidelines by analyzing the impact (namely, sensitivity) of each specific component on the LLMs: if a slight change in terms of a component could make a huge difference in LLMs's performance, then this component is significant as an guideline term. For example, using metaphors for emphasis instead of straightforward statements intended to better describe a task might lead to devastating performance cliff. Hence the guideline should take metaphors into account. Note that our analysis is centered around the zero-shot setting as the few-shot setting is already well-studied by prior arts (Brown et al., 2020; Lu et al., 2022) and the exemplars are not our focus. In this experiment and others in the following, we use *text-davinci-002* of GPT-3 in the zero-shot setting and set the temperature as zero (See Sec. B for more hyper-parameters). For all tasks and datasets, we follow the default setting of (Kojima et al., 2022) and run the experiments on test sets.

### 2.1 Impact of Syntax

We study the impact of syntax on the zero-shot prompting performance by comparing results from prompts of natural English and ones where word order is perturbed and grammatical errors are introduced.

#### 2.1.1 Grammatical Error Injection

**Setup** We produce perturbed prompts by artificially injecting grammatical errors into the question. Specifically, we study the problem on the StrategyQA task (Geva et al., 2021). The original question, as well as the options, if any, are used as the prompt. We additionally append "Therefore, the answer is:" to the prompt to induce question answering behaviors in LLMs. Without further notice, we use the same way to create original prompts for other factors. The perturbed prompt is generated by adding rule-based noises (Bryant et al., 2022) to the question and pairing it with the original options: see Tab. 1 for a complete list of the rule-based noises we use.

| Error Type | Error Description |
|---|---|
| Omi_Prep | Omission of the preposition. |
| Omi_Conj | Omission of the conjunction. |
| Omi_And | Using two consecutive adverbs of question or pronouns without "and". |
| Double_Noun | Double nouns. |
| Repeat_Word | Repeat a word. |
| Plu_Aft_Card | Using plural noun after cardinal numbers. |
| Super_Bef_Prep | Using a superlative adjective before preposition "than". |

Table 1: The list of the rule-based errors injected in the question.

**Results** The results in Tab. 2 show the accuracy of original prompts *vs*. perturbed prompts. We observe that grammatical errors do not necessarily have a negative impact on the zero-shot prompting. The robustness of GPT-3 against minor grammatical errors is similar to that of humans. For humans, minor grammatical errors that do not change semantics will not affect understanding either.

#### 2.1.2 Word Permutation

**Setup** While small perturbations like grammatical errors will not affect the performance, we further study severe perturbation to syntax: eliminating the syntactic structure by randomly shuffling words in a sentence (Kramsch, 2014). We experiment on two tasks: MultiArith (Roy and Roth, 2016) and StrategyQA (Geva et al., 2021). In the perturbed prompts, words are randomly reordered.

CCL 2024

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1408-1426, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

| Error Type | Original | Perturbed |
|---|---|---|
| Omi_Prep | 0.570 | 0.577 |
| Omi_Conj | 0.505 | 0.473 |
| Omi_And | 0.580 | 0.590 |
| Double_Noun | 0.537 | 0.553 |
| Repeat_Word | 0.540 | 0.530 |
| Plu_Aft_Card | 0.533 | 0.500 |
| Super_Bef_Prep | 0.286 | 0.429 |

Table 2: Comparison of accuracy on StrategyQA from original prompts and perturbed prompts on grammatical errors from Tab. 1.

**Results** The results are shown in Tab. 3. We observe that shuffling the words brings a notable decrease in zero-shot performance. It clearly illustrates that word order in the prompts affects the performance of LLMs. In summary, the experiments show that prompts should still follow the common word order while minor grammatical errors are negligible.

| | Original | Perturbed |
|---|---|---|
| MultiArith | 0.162 | 0.035 |
| StrategyQA | 0.593 | 0.473 |

Table 3: Comparison of accuracy on MultiArith and StrategyQA from original prompts and perturbed prompts on word order permutation.

**Few-shot Setting** Additionally, we study the effect of syntax on the few-shot setting, feeding GPT-3 with eight examples and the query. The word order of the questions in examples is shuffled for the perturbed prompts. See Tab. 18 for the results.

| | Origin | Perturbed |
|---|---|---|
| MultiArith | 0.327 | 0.305 |
| StrategyQA | 0.614 | 0.600 |

Table 4: Comparison of accuracy of original prompts *vs*. perturbed prompts with perturbations to syntax (Few-Shot Setting).

CCL 2024

### 2.2 Impact of Semantics

To what extent can LLMs understand a sentence when a negated word is injected to reverse the semantic meaning? We find the answer based on the comparison between prompts with and without negation.

**Setup** We experiment on three types of tasks and nine datasets constructed for experiments on semantic negation following (Jang et al., 2022): commonsense reasoning tasks on PIQA (Bisk et al., 2020), ARC-Easy (Clark et al., 2018), and COPA (Gordon et al., 2012), sentence completion tasks on HellaSwag (Zellers et al., 2019), StoryCloze (Mostafazadeh et al., 2016), and Lambada (Paperno et al., 2016), and question answering tasks on WQ (Berant et al., 2013), NQ (Kwiatkowski et al., 2019), and TriviaQA (Joshi et al., 2017). All sample sizes are 300 except for the COPA dataset, which has a sample size of 100. For each dataset, we compare the original prompts with the perturbed prompts. For the perturbed prompt, we change the question to be expressed using negative phrases. As there are two questions, the originally correct choice will be wrong while the originally wrong one will turn correct.

**Results** From the results shown in Tab. 5, we find that negation brings continued challenges for LLMs on various tasks, and yet human beings handle it very well (Jang et al., 2022). The results suggest that when writing a prompt, users should avoid using negative words.It's better to be clear without ambiguity, and show obvious and concrete goals.

| | Original | Perturbed |
|---|---|---|
| ARC-Easy | 0.960 | 0.907 |
| PIQA | 0.840 | 0.783 |
| COPA | 0.970 | 0.520 |
| HellaSwag | 0.663 | 0.473 |
| StoryCloze | 0.933 | 0.630 |
| Lambada | 0.593 | 0.440 |
| WQ | 0.933 | 0.803 |
| TriviaQA | 0.963 | 0.973 |
| NQ | 0.910 | 0.880 |

Table 5: Comparison of accuracy on various datasets from original prompts and perturbed prompts for semantics.

### 2.3 Impact of Lexicon

Here we focus on which word to choose when multiple words with closely related meanings exist. We approach the problem by comparing the original prompts with perturbed prompts where certain words are replaced with their synonyms or translations, *e.g.*, Chinese, while keeping the semantic meaning.

#### 2.3.1 Synonym Substitution

**Setup** Experiments are conducted on two tasks: HellaSwag (Zellers et al., 2019) and COPA (Gordon et al., 2012). For the question of a perturbed prompt, we randomly pick up a notional word and replace it with its synonym based on WordNet (Bird et al., 2009). Specifically, the question is tagged using a part-of-speech tagger (Bird et al., 2009) to find out a notional word such as a noun or verb. Then the notional word is substituted with its synonym from WordNet (Bird et al., 2009). We only change one notional word per prompt. If no words in a prompt have synonyms in WordNet, we will keep the prompt as is. To make sure the semantics of the prompt is preserved after the word substitution, BERTScore (F1-score) (Zhang et al., 2019) between the original prompt and the perturbed prompt is calculated to filter out those with low scores. Options remain unchanged.

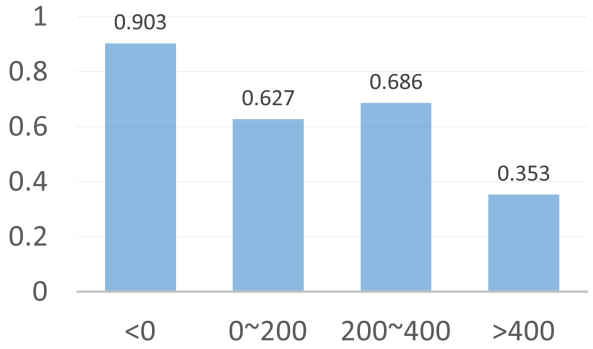



Figure 2: Relationship between the familiarity degree and performance decrease. The experiment is conducted on *HellaSwag*. The x-axis refers to the familiarity score of the original prompt minus the familiarity score of the perturbed prompt. The larger this difference is, the greater the unfamiliarity. The y-axis represents the performance of the perturbed prompt.

CCL 2024

**Results** The experimental results are shown in Tab. 6. Compared with the original prompts, the perturbed prompts lead to inferior performance. Furthermore, we investigate the reasons behind the degradation. We mask the replaced word and calculate the conditional probability of the word masked to measure familiarity and denote it's log value as the familiarity score. As shown in Fig. 2, performance is positively correlated with familiarity. The more familiar the prompts, the better the model performs.

| | **Original** | **Perturbed** |
|---|---|---|
| HellaSwag | 0.60 | 0.55 |
| COPA | 0.97 | 0.95 |

Table 6: Comparison of accuracy on HellaSwag and COPA from original prompts and perturbed prompts on synonym substitution.

### 2.3.2 Translation Substitution

**Setup** We further experiment with word substitution with another language. We evaluate translation substitution on a randomly selected set of 300 samples from StrategyQA training dataset. Similar to the word substitution, the perturbed question substitutes a word with its translation in another language. We study English to Chinese translation in this work.

**Results** The results are shown in Tab. 7. In general, the two substitution methods achieve relatively the same performance compared with the original prompting method. However, translation substitution has a bigger impact than synonym substitution. This is predictable as the perplexity score of the perturbed question with a synonym is higher than the perturbed question with a translated word in another language. It's suggested that users should choose words that are more statistically prevalent in the training corpus of LLMs.

| | **Original** | **Perturbed** |
|---|---|---|
| Synonym | 0.540 | 0.541 |
| Translation | 0.540 | 0.530 |

Table 7: Comparison of accuracy of original prompts *vs*. perturbed prompts with changes on lexicon.

**Zero-Shot CoT** We also conduct the experiment to study lexicon's effect on the zero-shot CoT via translation substitution. The results are shown in Tab. 19.

| | **Original** | **Perturbed** |
|---|---|---|
| Zero-Shot | 0.540 | 0.530 |
| Zero-Shot CoT | 0.513 | 0.497 |

Table 8: Comparison of accruacy of original prompts *vs*. perturbed prompts with perturbations to the lexicon.

## 2.4 Impact of Pragmatics

We primarily test metaphor understanding and compare original prompts with and without metaphors.

**Setup** For experiments on pragmatics, we design a new task based on the Emotion Recognition dataset on SemEval2018 (Mohammad et al., 2018) and TweetEval (Barbieri et al., 2020). We name the task Prompt-based Metaphor Emotion. The task is a multi-choice classification task and aims to recognize four basic emotional expressions: anger, joy, sadness, and optimism. A model is expected to pick one according to the question. We design two types of questions, with and without metaphors. The questions including original prompts and eight perturbed sets are

CCL 2024

shown in Tab. 9. The entire dataset contains 1421 samples, covering 558, 358, 382, and 123 samples for anger, joy, sadness, and optimism, respectively.

| *Prompt Template for Analysis of Pragmatics* |
|---|
| Please choose the appropriate answer for the following question from A, B, C, D.<br>Which of the following sentences expresses the emotion like "*¡Emotion Description¿*"?<br>A: You know, she is a disabled figure. But why did you still laugh at her? Do you want to push her buttons?<br>B: The film tells the story of Miguel, a little boy from the shoemaker's family. His family has a mysterious ban —-— No one in the family can touch music performances.<br>C: When the adorable kid met her favorite idol, she was suddenly weeping with joy.<br>D: On the Day of the Dead in Mexico, the dead return to the real world and gather with their loved ones. Leave me alone, ok? I'm in a bad mood now. |

*¡Emotion Description¿*

| **ID** | Emotion of Anger | Emotion of Joy | Emotion of Sadness | Emotion of Optimism |
|---|---|---|---|---|
| **Original** | Anger | Joy | Sadness | Optimism |
| **Perturbed_1** | Flew off the handle | Jumping up and down with excitement | Down in the dumps | Every cloud has a silver lining |
| **Perturbed_2** | Lose your cool | Grinning from ear to ear | Feel blue | When one door closes, another one opens |
| **Perturbed_3** | Blow a fuse | In seventh heaven | Heart is broken | There is light at the end of the tunnel |
| **Perturbed_4** | Go ballistic | Walking on air | Reduced to tears | When life hands you lemons, make lemonade |
| **Perturbed_5** | Hit the ceiling | On cloud nine | Cry one's eyes out | Count your blessings |
| **Perturbed_6** | Volcanic eruption | Over the moon | Lump in one's throat | Hope against hope |
| **Perturbed_7** | Flip one's wig | On top of the world | Under the weather | There are plenty of fish in the sea |
| **Perturbed_8** | Raise the roof | A pig in mud | Get the face of a wet weekend | See the world through rose-colored glasses |

Table 9: The eight perturbed sets of prompts used to express anger, joy, sadness and optimism, respectively. The same column indicates the same emotion.

**Results** The results are shown in Tab. 10. The perturbed prompts generally underperform the straightforward ones in each emotion. Therefore, we recommend not using metaphors, but stating ones' needs directly. If metaphors are necessary, it is recommended to choose those that are very common and widely used, such as *Every cloud has a silver lining*.

## 3 `Prompt Engineering 101`

Since Sec. 2 show clear paths for us to navigate in the space of prompt design, in this section, we discuss good practices for an amateur to interact with LLMs via prompting, following the analytical results aforementioned. In addition to general requirements, *e.g.*, the language choice, human readability, avoiding abstract statement (Khashabi et al., 2022), prompt engineering should follow four main guidelines as `Prompt Engineering 101`:

CCL 2024

| ID | Anger | Joy | Sadness | Optimism |
|---|---|---|---|---|
| **Original** | 0.772 | 0.688 | 0.696 | 0.538 |
| **Perturbed_1** | 0.668 | 0.576 | 0.626 | 0.544 |
| **Perturbed_2** | 0.626 | 0.562 | 0.608 | 0.532 |
| **Perturbed_3** | 0.602 | 0.516 | 0.582 | 0.516 |
| **Perturbed_4** | 0.578 | 0.504 | 0.582 | 0.490 |
| **Perturbed_5** | 0.554 | 0.502 | 0.522 | 0.478 |
| **Perturbed_6** | 0.444 | 0.496 | 0.484 | 0.422 |
| **Perturbed_7** | 0.376 | 0.474 | 0.458 | 0.392 |
| **Perturbed_8** | 0.194 | 0.330 | 0.432 | 0.388 |

Table 10: Comparison of accuracy of original prompts and perturbed prompts on metaphoric pragmatics.

- Checking Syntax: Sentences should be free from grammatical errors and respect proper word order. If possible, a grammatical error checker will be helpful for users to check their language.
- Checking Semantics: Sentences should be clear without ambiguity. Users should describe their goals and intentions directly without using many negations.
- Checking Lexicon: Without changing the meaning, a user should choose words that are prevalent for training LLMs.
- Checking Pragmatics: We recommend not using metaphors, but stating the needs straightforwardly. If metaphors are necessary, it is recommended to choose those that are very common and widely used.

Fig. 1 illustrates the flow diagram of our proposed guidelines. It is a pipeline including four stages. Our prescribed approach first requires grammatical error correction to check the input sentence. With grammar errors free, the prompt will further go through analysis on semantics, removing unnecessary negations, common word substitution, and metaphor removal.

**Guidelines Automation** We further automated `Prompt Engineering 101` to build an **automatic toolkit** designed based on our guidelines. The specific automation implementation program is as follows. First, we performed spelling corrector [0] and grammar checker [1] on the original prompt. Then, we used ChatGPT [2] for synonym replacement following negation detection [3]. If negations exist, a rephrasing tool was used to rephrase the sentence. Finally, we performed metaphor detection [4]. Prompts designed for ChatGPT is shown in Tab. 11. We used the OpenAI official interface for the experiments, modeled as `gpt-3.5-turbo` [5].

## 4 Validation of `Prompt Engineering 101`

### 4.1 Setup

We validate the proposed guidelines with 7 selected tasks from BIG-bench Lite (BBL) (Srivastava et al., 2022): Auto Debugging, Code Line Description, Formal Fallacies Syllogisms Negation, Hindu Knowledge, Operators, Play Dialog Same or Different and Language Identification. Each of these tasks tests a specific aspect of the capabilities of language models. For each dataset, we choose at most 100 instances to conduct our experiment. If the size of the dataset is less than 100, we adopt the whole dataset for our experiment. The hyper-parameters are identical to the analysis experiments in Sec. 2. The experiments are conducted with ten amateurs over these seven tasks. We recruit human subjects from Prolific. More specifically, we create a survey

[0] https://github.com/Rudransh11Kohli/Grammarly
[1] https://github.com/bhattbhavesh91/gramformer-tutorial
[2] https://openai.com/blog/chatgpt
[3] https://github.com/evanmiltenburg/negation-detector
[4] https://huggingface.co/lwachowiak/Metaphor-Detection-XLMR
[5] Following paper (Kojima et al., 2022), temperature=0 and the other parameters are defaulted.

| Phase | Prompt |
|---|---|
| **Synonym** | `Please replace the sentence "<sentence>" with synonyms so that the semantics of the sentence remain unchanged. Only replace words, do not change the structure and length of sentences, and do not replace if there are no alternatives. Please output the modified sentence, do not output redundant content.` |
| **Negation** | `The following sentence "<sentence>" contains negation "<negation>", please rewrite the sentence so that it does not contain negation and keeps the semantics. Note that negation refers to a phrase or word with a negative meaning. Please output the modified sentence, do not output redundant content.` |
| **Metaphor** | `The following sentence "<sentence>" contains metaphor, please rewrite the sentence so that it does not contain metaphor and keeps the semantics. Please output the modified sentence, do not output redundant content.` |

Table 11: Prompts of ChatGPT.

that allows workers to understand what the task is, given labeled instances as well as the task title. We first provide workers with comprehensive instructions and a set of unlabeled instances to judge if a worker understands the task. During the survey, we show workers an interface for writing down their descriptions of a given task as the prompts. After the initial prompts[6] are written, we recruit another ten workers to refine the initial prompts with and without our guidelines (namely, guidelines for human or automatic toolkit). We compare the performance of the amateurs' initial prompts and the new set of prompts refined with/without the help of our guidelines on the randomly sampled tasks and test instances.[7]

### 4.2 Results With *vs*. Without Guidelines for Humans

As evidenced by (Mishra et al., 2022; Efrat and Levy, 2020), original prompts (**Row** `initial prompts`) written by amateurs are not easy to follow for LLMs. The same observation can also be found in our experiments. All results are shown in Tab. 12. After humans refine the prompts without the help of our guidelines (**Row** `w/o human`), the performance increase from 54.73% to 57.04%. The guidelines for humans further increase it to 57.46% (**Row** `w human`).

### 4.3 Results With *vs*. Without Guideline for Automatic Toolkit

As shown in Tab. 12, the average performance improvement achieved by applying our automated toolkits (**Row** `w automatic`) is 1.31% (from 54.73% to 56.04%). More importantly, we further combine guidelines for humans and automatic toolkit to refine the prompts and got a improvement of 3.37% (**Row** `w human + automatic`).

| | **Perf.** | △ |
|---|---|---|
| Initial prompts | 54.73 | - |
| w automatic | 56.04 | +1.31 |
| w/o human | 57.04 | +2.31 |
| w human | 57.46 | +2.73 |
| w human + automatic | **58.10** | +3.37 |

Table 12: Results with/without our guidelines for humans and automatic toolkit. Perf.: Performance.

### 4.4 Ablation Study

We conduct an ablation study for Sec. 4.2 to reveal the functions of each checker w.r.t each linguistic factor. As shown in Tab. 13, all checkers play an important role in improving writing

[6]There are 70 initial prompts and about 7000 instances in total.
[7]Performance in this section refers to accuracy.

Proceedings of the 23rd China National Conference on Computational Linguistics, pages 1408-1426, Taiyuan, China, July 25 - 28, 2024.
Volume 1: Main Conference Papers
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

prompts. They complement each other and contribute to the final improvement. Most errors occur in the stage of lexicon checking, indicating that many prompt writers are unaware of using more common words for LLMs. Our guidelines provide useful advice in this regard. Sometimes, writers are careless with their spelling, word order, and the use of ambiguous words and negation, which should be eliminated according to our guidelines.

| | #Mod. | Perf. |
|---|---|---|
| `Prompt Engineering 101` | 59 | 57.46 |
| -Pragmatic Checker | 58 | 57.36 |
| -Lexical Checker | 24 | 54.77 |
| -Semantic Checker | 15 | 54.76 |
| -Syntactic Checker | 0 | 54.73 |

Table 13: #Mod.: The number of modifications for each checker. Perf.: Task performance.

### 4.5 Significance Analysis

We use the accuracy of the 10 experiments (each set contains initial prompts and refined prompts). The null hypothesis (H0) is that the guidelines are beneficial to the performance. The significance value is 0.05. We conducted the two-sided t-test. The calculated t-value is -2.30, and the corresponding p-value is 0.034, lower than the selected significance value of 0.05. Therefore, the null hypothesis holds, which means the guidelines are helpful for improving the performance of GPT-3.

### 4.6 Model Generalization Validation

To verify whether our method has model generalization, we perform negation experiments on four different GPT-3 variants of different sizes (`text-ada-001`, `text-babbage-001`, `text-curie-001` and `text-davinci-003`). The results are shown in Tab. 20. We found that the experimental group with negations generally performed worse than the control group. This conclusion is consistent with the experiments on `text-davinci-002`, so we consider our approach to be model generalizable.

| | ARC-Easy | PIQA | COPA | HellaSwag | StoryCloze | Lambada | WQ | TriviaQA | NQ |
|---|---|---|---|---|---|---|---|---|---|
| **text-ada-001**(1.3B) | 0.427/**0.467** | **0.533**/0.407 | 0.400/**0.440** | 0.380/**0.400** | 0.443/**0.497** | 0.247/**0.323** | 0.393/**0.453** | 0.410/**0.473** | 0.403/**0.557** |
| **text-babbage-001**(6.7B) | 0.463/**0.527** | 0.473/**0.523** | **0.500**/0.470 | 0.197/**0.240** | 0.483/**0.497** | **0.500**/0.493 | 0.477/**0.503** | 0.480/**0.523** | 0.483/**0.517** |
| **text-curie-001**(13B) | 0.497/**0.540** | 0.480/**0.523** | 0.090/**0.290** | 0.213/**0.337** | 0.490/**0.503** | **0.497**/0.460 | 0.473/**0.540** | 0.453/**0.567** | 0.470/**0.540** |
| **text-davinci-003**(175B) | 0.820/**0.963** | 0.763/**0.833** | 0.420/**0.960** | 0.267/**0.433** | 0.677/**0.907** | 0.287/**0.680** | **0.977**/0.967 | **0.980**/0.978 | 0.877/**0.943** |

Table 14: Negations experiments on four GPT-3 variants of different sizes. Note that metrics representing **neg/pos**.

## 5 Conclusion and Future Work

We investigate and analyze the sensitivity of prompts from the perspective of linguistics, covering four factors, *i.e.*, syntax, semantics, lexicon, and pragmatics, and their impact on various tasks. Based on the results, we streamline a list of guidelines as `Prompt Engineering 101` for prompt engineering. Via human evaluation, we show the effectiveness of the proposed guidelines. We further automated `Prompt Engineering 101` to build the automatic toolkit designed based on our guidelines. After replacing manual refinement with an automatic toolkit, we still observed an average improvement in performance. This further enhances the superiority of `Prompt Engineering 101`. There are more linguistic factors, for example, phonetics, phonology and typology, for us to study. Even for syntax, there remain aspects that we have not explored. Yet, we take the first step to exploring four of them. A more in-depth analysis for other factors on a wider range of tasks should be favorable. This work also provides another window into prompt engineering that regards prompts as a specific "langauge" to interact with LLMs. This

CCL 2024

"language", like a dialect, is a variety of language that is characteristic of LLMs. From the perspective of behaviors, they have their own grammatical and phonological rules, linguistic features, and stylistic aspects, that is potentially different from the natural language. With mounting evidence, LLMs use prompts in a different way from humans. Nevertheless, there is no clear consensus on what kind of linguistic factor plays the most important role. On top of that, we hope that this work would call for future research into the intriguing problem.

## 6 Acknowledgements

This work is supported by the Talent Fund of Beijing Jiaotong University (2023XKRC006). The authors would like to thank the anonymous reviewers for their valuable comments and suggestions to improve this paper.

## References

Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. 2020. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. *arXiv preprint arXiv:2010.12421*.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1533–1544.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.

Gwern Branwen. 2020. Gpt-3 creative fiction. *The Website of Gwern Branwen*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Christopher Bryant, Zheng Yuan, Muhammad Reza Qorib, Hannan Cao, Hwee Tou Ng, and Ted Briscoe. 2022. Grammatical error correction: A survey of the state of the art. *arXiv preprint arXiv:2211.05166*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Avia Efrat and Omer Levy. 2020. The turking test: Can language models understand instructions? *arXiv preprint arXiv:2010.11982*.

Dana Ferris. 1999. The case for grammar correction in l2 writing classes: A response to truscott (1996). *Journal of second language writing*, 8(1):1–11.

Daniel Fried, Nicholas Tomlin, Jennifer Hu, Roma Patel, and Aida Nematzadeh. 2022. Pragmatics in grounded language learning: Phenomena, tasks, and modeling approaches. *arXiv preprint arXiv:2211.08371*.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Talmy Givón. 1978. Negation in language: pragmatics, function, ontology. In *Pragmatics*, pages 69–112. Brill.

CCL 2024

Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2012. Semeval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 394–398.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2022. A fine-grained comparison of pragmatic language understanding in humans and language models. *arXiv preprint arXiv:2212.06801*.

Joel Jang, Seonghyeon Ye, and Minjoon Seo. 2022. Can large language models truly understand prompts? a case study with negated prompts. *arXiv preprint arXiv:2209.12711*.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know. *Transactions of the Association for Computational Linguistics*, 8:423–438.

Mark D Johnson, Anthony Acevedo, and Leonardo Mercado. 2016. Vocabulary knowledge and vocabulary use in second language writing. *TESOL Journal*, 7(3):700–715.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Dilek Karakoç and Gül Durmuşoğlu Köse. 2017. The impact of vocabulary knowledge on reading, writing and proficiency scores of efl learners. *Journal of language and linguistic studies*, 13(1):352–378.

Gabriele Kasper and Carsten Roever. 2005. Pragmatics in second language learning. *Handbook of research in second language teaching and learning*, pages 317–334.

Daniel Khashabi, Chitta Baral, Yejin Choi, and Hannaneh Hajishirzi. 2022. Reframing instructional prompts to gptk's language. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 589–612.

Junyeob Kim, Hyuhng Joon Kim, Hyunsoo Cho, Hwiyeol Jo, Sang-Woo Lee, Sang-goo Lee, Kang Min Yoo, and Taeuk Kim. 2022. Ground-truth labels matter: A deeper look into input-label demonstrations. *arXiv preprint arXiv:2205.12685*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*.

Kepa Korta and John Perry. 2020. Pragmatics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Spring 2020 edition.

Claire Kramsch. 2014. Language and culture. *AILA review*, 27(1):30–55.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland, May. Association for Computational Linguistics.

John Lyons. 1981. *Language and linguistics*. Cambridge university press.

Aman Madaan and Amir Yazdanbakhsh. 2022. Text and patterns: For effective chain of thought, it takes two to tango. *arXiv preprint arXiv:2209.07686*.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487, Dublin, Ireland, May. Association for Computational Linguistics.

CCL 2024

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 1–17.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849.

Jonas Oppenlaender. 2022. A taxonomy of prompt modifiers for text-to-image generation. *arXiv preprint arXiv:2204.13988*.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The lambada dataset: Word prediction requiring a broad discourse context. *arXiv preprint arXiv:1606.06031*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. 2022. Impact of pretraining term frequencies on few-shot reasoning. *arXiv preprint arXiv:2202.07206*.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7.

Subhro Roy and Dan Roth. 2016. Solving general arithmetic word problems. *arXiv preprint arXiv:1608.01413*.

Tony Silva and Paul Kei Matsuda. 2012. *On second language writing*. Routledge.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Albert Webson and Ellie Pavlick. 2021. Do prompt-based models really understand the meaning of their prompts? *arXiv preprint arXiv:2109.01247*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, et al. 2023. Zero-shot information extraction via chatting with chatgpt. *arXiv preprint arXiv:2302.10205*.

Xiang Wei, Yufeng Chen, Ning Cheng, Xingyu Cui, Jinan Xu, and Wenjuan Han. 2024. CollabKG: A learnable human-machine-cooperative information extraction toolkit for (event) knowledge graph construction. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3490–3506, Torino, Italia, May. ELRA and ICCL.

Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In *CHI Conference on Human Factors in Computing Systems*, pages 1–22.

CCL 2024

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Ningyu Zhang, Luoqiu Li, Xiang Chen, Shumin Deng, Zhen Bi, Chuanqi Tan, Fei Huang, and Huajun Chen. 2021. Differentiable prompt makes pre-trained language models better few-shot learners. In *International Conference on Learning Representations*.

## A Related Work

Successful practices usually require domain expertise and are extremely time-consuming—one needs to spend a significant amount of time on wording since a slight change in the textual input could have a huge impact on performance (Reynolds and McDonell, 2021). As a result, "engineering" these prompts has been one of the major challenges in deploying LLMs in real-world applications. (Webson and Pavlick, 2021). Using the recently introduced In-Context Learning (ICL), a PLM only needs to take in a few training examples and a test instance as context and directly decode the output without any updates to its parameters. However, using open-ended natural language as contextual instruction in ICL is a double-edged sword; while users can input anything and access a variety of generated sentences, they also must engage in cumbersome trials and errors when the resulting language quality is poor (Jiang et al., 2020; Liu et al., 2021). Successful practices usually require domain expertise and are extremely time-consuming—one needs to spend a significant amount of time on wording since a slight change in the textual input could have a huge impact on performance (Reynolds and McDonell, 2021). As a result, "engineering" these prompts has been one of the major challenges in deploying LLMs in real-world applications. (Webson and Pavlick, 2021). The study of prompt engineering can be traced back to T5 (Raffel et al., 2020) and GPT-2 (Radford et al., 2019), where single word task identifiers were utilized to distinguish between different tasks. The term "prompt engineering" was originally coined by (Branwen, 2020) when describing GPT-3's few-shot ICL capabilities. (Branwen, 2020) positioned prompting as an emerging technique for LLMs that were trained from Internet-scale data (Wu et al., 2022). It's recently shown that prompting is an effective method to elicit specific knowledge and abstractions needed to perform well on unseen tasks and bears potential to help end users interact with intelligent language models. Demonstrations in the prompt are well-studied by prior arts. (Lu et al., 2022) and (Kim et al., 2022) reveal the role of providing format clues, the sensitivity of the order of examples, and the importance of ground-truth labels, respectively. This work and many others along the same vein established foundation within the community to study the function of different language aspects in prompting. (Madaan and Yazdanbakhsh, 2022) systematically experiment with symbols, patterns, and text to study these items separately. Another line of research is Chain-of-Thought (CoT) designing (Wei et al., 2022). A manually designed rationale benefit the prompting performance. (Khashabi et al., 2022) share a similar basis with CoT (Wei et al., 2022). They claim that manually decomposing a task instruction into multiple simpler subtasks leads to better performance. Meanwhile, hobbyists and practitioners on the Internet have already engaged in prompt engineering. They discuss fantastic "tricks" and keywords and discover how to steer LLMs toward desirable aesthetic generations (Oppenlaender, 2022). For example, using "unreal engine" as a prompt will help add a hyper-realistic, 3D render quality to the image generation [8].

## B Hyper-parameters

Hyper-parameters of the GPT-3 are listed in Tab. 15.

## C Data Statistic

We show the data statistics for semantics analysis in Tab. 16 and lexicon analysis in Tab. 17.

[8]Aran Komatsuzaki's findings at https://twitter.com/arankomatsuzaki/status/1399471244760649729.

| Hyper-parameter | Value |
|---|---|
| Max token | 64 |
| Top-p | 1 |
| Temperature | 0 |
| Frequency Penalty | 0 |
| Presence Penalty | 0 |

Table 15: Hyper-parameters of the GPT-3 model we use.

| Dataset | #Instances |
|---|---|
| ARC-Easy | 300 |
| PIQA | 300 |
| COPA | 100 |
| HellaSwag | 300 |
| StoryCloze | 300 |
| Lambada | 300 |
| WQ | 300 |
| TriviaQA | 300 |
| NQ | 300 |

Table 16: Data statistic of experiments for impact analysis of semantics.

| Dataset | #Instances |
|---|---|
| PIQA | 50 |
| ARC-Easy | 50 |
| HellaSwag | 50 |
| HellaSwag | 300 |
| SuperGlue | 50 |
| SuperGlue | 100 |

Table 17: Data statistic of experiments for impact analysis of lexicon.

## D Few-shot Setting for Syntax

Additionally, we study the effect of syntax on the few-shot setting, feeding GPT-3 with eight examples and the query. The word order of the questions in examples is shuffled for the perturbed prompts. See Tab. 18 for the results.

| | MultiArith | StrategyQA |
|---|---|---|
| *Few-Shot Setting* | | |
| **Original** | 32.7 | 61.4 |
| **Perturbed** | 30.5 | 60.0 |
| *Zero-Shot Setting* | | |
| **Original** | 16.2 | 59.3 |
| **Perturbed** | 3.5 | 47.3 |

Table 18: Comparison of original prompts *vs*. perturbed prompts with perturbations to syntax.

CCL 2024

## E Zero-Shot CoT for Lexicon

We also conduct the experiment to study lexicon's effect on the zero-shot CoT via translation substitution. The results are shown in Tab. 19.

| | Zero-Shot | Zero-Shot CoT |
|---|---|---|
| **Original** | 54.0 | 51.3 |
| **Perturbed** | 53.0 | 49.7 |

Table 19: Comparison of original prompts *vs*. perturbed prompts with perturbations to the lexicon. We show the effect via translation substitution.

## F Model Generalization Validation

To verify whether our method has model generalization, we perform negation experiments on four different GPT-3 variants of different sizes (`text-ada-001`, `text-babbage-001`, `text-curie-001` and `text-davinci-003`). The results are shown in Tab. 20. We could find that the experimental group with negations generally performed worse than the control group. This conclusion is consistent with the experiments on `text-davinci-002`, so we consider our approach to be model generalizable.

| | ARC-Easy | PIQA | COPA | HellaSwag | StoryCloze | Lambada | WQ | TriviaQA | NQ |
|---|---|---|---|---|---|---|---|---|---|
| **text-ada-001**(1.3B) | 0.427/**0.467** | **0.533**/0.407 | 0.400/**0.440** | 0.380/**0.400** | 0.443/**0.497** | 0.247/**0.323** | 0.393/**0.453** | 0.410/**0.473** | 0.403/**0.557** |
| **text-babbage-001**(6.7B) | 0.463/**0.527** | 0.473/**0.523** | **0.500**/0.470 | 0.197/**0.240** | 0.483/**0.497** | **0.500**/0.493 | 0.477/**0.503** | 0.480/**0.523** | 0.483/**0.517** |
| **text-curie-001**(13B) | 0.497/**0.540** | 0.480/**0.523** | 0.090/**0.290** | 0.213/**0.337** | 0.490/**0.503** | **0.497**/0.460 | 0.473/**0.540** | 0.453/**0.567** | 0.470/**0.540** |
| **text-davinci-003**(175B) | 0.820/**0.963** | 0.763/**0.833** | 0.420/**0.960** | 0.267/**0.433** | 0.677/**0.907** | 0.287/**0.680** | **0.977**/0.967 | **0.980**/0.978 | 0.877/**0.943** |

Table 20: Negations experiments on four GPT-3 variants of different sizes. Note that metrics representing **neg/pos**.

## G Results details

Tab. 21 provides the detailed results obtained using either our guidelines for human or automatic toolkit. The metrics used are micro accuracy and macro accuracy.

- **Micro accuracy** measures the probability of correctly identifying all samples across the seven tasks.
- **Macro accuracy** calculates the accuracy for each of the seven tasks and then averages them.

To obtain the final performance w.r.t. guidelines for humans, we calculate the accuracy for each human and then average their scores.

| | micro acc | macro acc |
|---|---|---|
| baseline | 54.73 | 54.57 |
| w auto | 56.04 | 55.23 |
| w/o human guidelines | 57.04 | 56.13 |
| w human guidelines | 57.46 | 56.54 |
| w human and auto guidelines | **58.10** | **57.24** |

Table 21: Detailed results with/without our human/auto guidelines.

CCL 2024

## H Human Study Details

Firstly, we released a questionnaire (See Fig. 3 for details) to collect prompts without our guidelines from 10 persons. At the beginning of the questionnaire, there is a general description and two examples of writing prompts. After that, we show the brief introduction of each task. After each introduction we provide 3 examples (the input and the expected output of the models) to help subjects better comprehend what the task aims at. At the end of each task we remind the subjects to write a general instruction of the task rather than a specific prompt based on the examples we provided. The subjects were required to write a prompt of at least 10 characters for each task. Finally we collected 10 valid results, and each result includes 7 valid prompts for all tasks.

Secondly, in order to obtain the prompts with our guidelines, we released another 10 questionnaires (see Fig. 4 for details) to collect refined prompts. The setting of the questionnaire is similar with the questionnaire we designed to collect baseline prompts, except for two exceptions. First, we modified the description and the examples in the beginning in order to collect refined prompts. Second, after the introduction and examples of each task, we also provided the original prompt and our guidelines to make subjects refine the prompts based on our guidelines. For each group of prompts written by the same person, we design a personalized questionnaire and ask one person to refine one group of prompts.

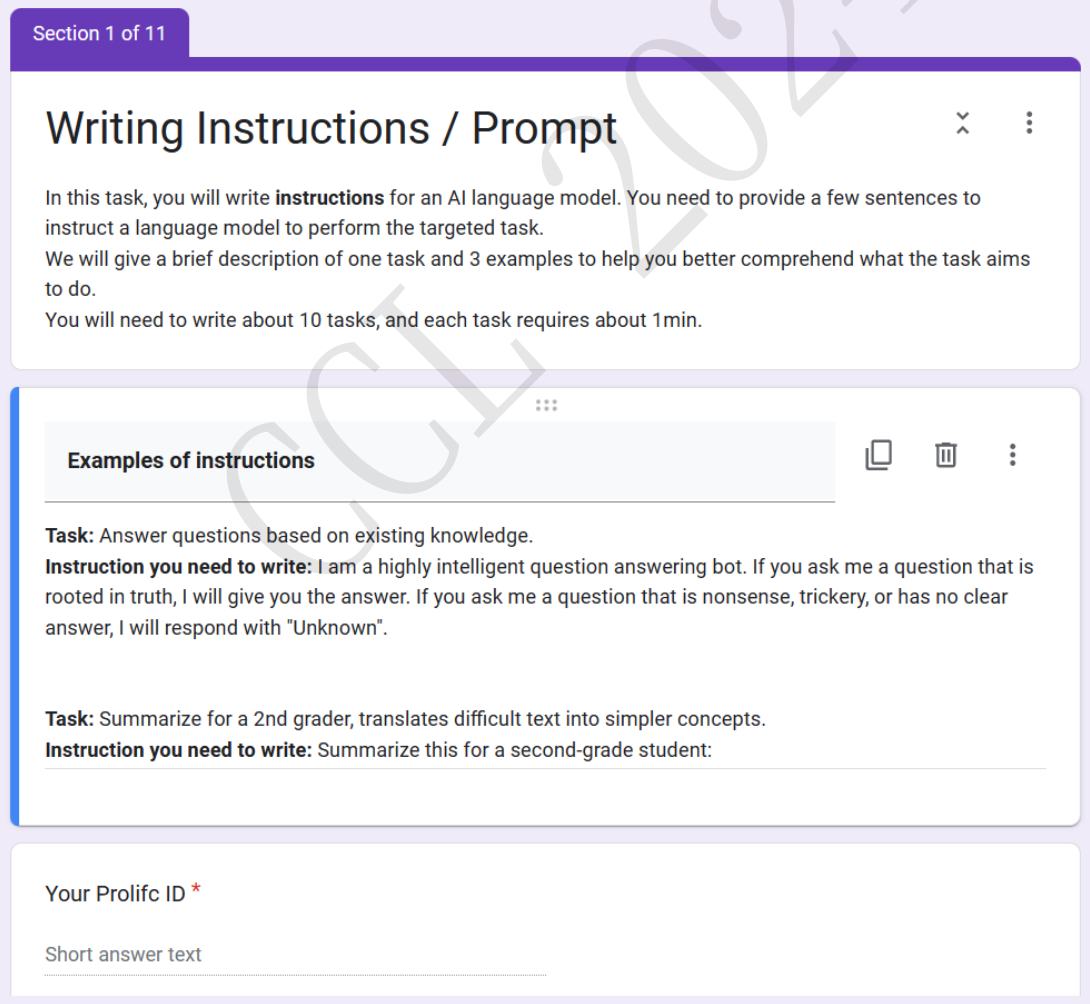



Figure 3: Interface of questionnaires without `Prompt Engineering 101`.

CCL 2024

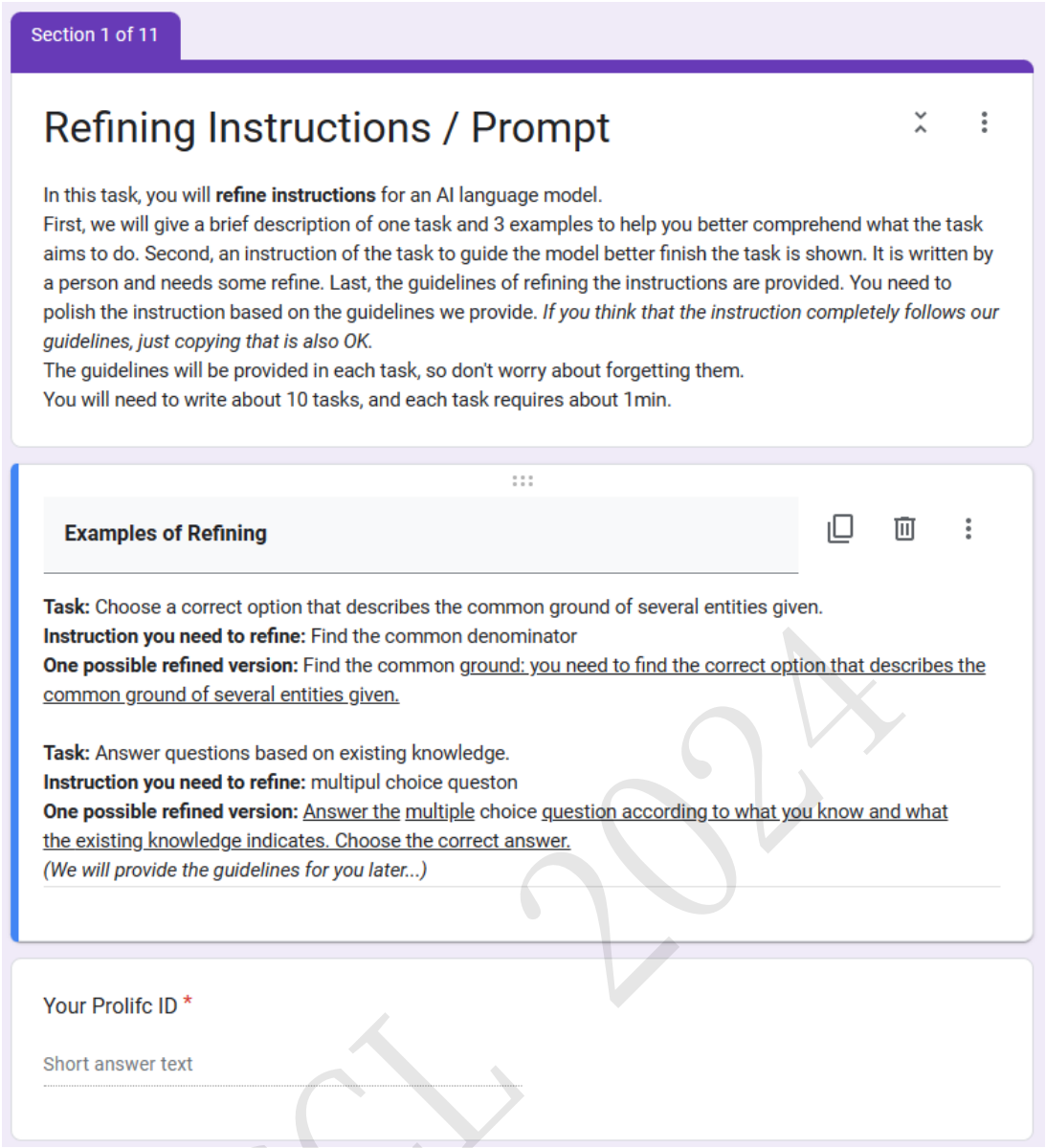



Figure 4: Interface of questionnaires with guidelines following `Prompt Engineering 101`.

## I Case Study

Examples of original prompts *vs*. perturbed prompts with changes on syntax, semantics, lexicon, and pragmatics are shown in Tab. 22.

CCL 2024

| | Original Prompt | Perturbed Prompt |
|---|---|---|
| Syntax | ***Prompt:*** For Halloween Deddy and her sister combined the candy they received. Deddy had 32 pieces of candy while her sister had 42. If they ate 35 pieces the first, how many pieces do they have left?<br>A: Therefore, the answer (arabic numberals) is | ***Prompt:*** Deddy the have the many they pieces Debby left? of combined her sister had how do candy Halloween candy while pieces If night, they 32 ate For 35 received. had 42. pieces and sister first her they<br>A: Therefore, the answer (arabic numberals) is |
| | ***Expected Output:*** 39 ***Output:*** 39 | ***Expected Output:*** 39 ***Output:*** 74 |
| Semantics | ***Prompt:*** Generate the correct answer to the following question.<br><br>Question: Earth, along with the other planets, revolves around the Sun. The planets revolve in a counterclockwise direction. The cause of the revolution is mostly due to which force?<br>A. Gravitational<br>B. Magnetic<br><br>Answer: | ***Prompt:*** Generate the incorrect answer to the following question.<br><br>Question: Earth, along with the other planets, revolves around the Sun. The planets revolve in a counterclockwise direction. The cause of the revolution is mostly due to which force?<br>A. Gravitational<br>B. Magnetic<br><br>Answer: |
| | ***Expected Output:*** A ***Output:*** A | ***Expected Output:*** B ***Output:*** A |
| Lexicon | ***Prompt:*** Are Christmas trees dissimilar to deciduous trees? | ***Prompt:*** Are Christmas 树 dissimilar to deciduous trees? |
| | ***Expected Output:*** Yes. ***Output:*** yes, Christmas trees are dissimilar to deciduous trees. | ***Expected Output:*** Yes. ***Output:*** no. |
| Pragmatics | ***Prompt:*** Please choose the appropriatee answer for the following question from A, B, C, D.<br><br>Which of the following sentences expresses the emotion like "anger"?<br>A. I've used almost half of my printing money and it's the first day of the semester. #pissed<br>B. When I chirp, shawty, chirp back.<br>C. U know what's very pathetic? The fact that I dearly miss my professors but they probably forgot about me already. #sad #:(<br>D. @user said, "Surely the #bitterness of #death is past." [2/2]<br><br>Answer: | ***Prompt:*** Please choose the appropriate answer for the following question from A, B, C, D.<br><br>Which of the following sentences expresses the emotion like "volcanic eruption"?<br>A. I've used almost half of my printing money and it's the first day of the semester. #pissed<br>B. When I chirp, shawty, chirp back.<br>C. U know what's very pathetic? The fact that I dearly miss my professors but they probably forgot about me already. #sad #:(<br>D. @user said, "Surely the #bitterness of #death is past." [2/2]<br><br>Answer: |
| | ***Expected Output:*** A ***Output:*** A | ***Expected Output:*** A ***Output:*** B |

Table 22: Examples of original prompts *vs*. perturbed prompts with changes on syntax, semantics, lexicon, and pragmatics. The text following ***Prompt*** represents the prompt. Original and perturbed prompts are denoted with red and green respectively.

CCL 2024